HUMBOLDT-UNIVERSITÄT ZU BERLIN

Development of cheminformatics-based methods for computational prediction of off-target activities

Dissertation
zur Erlangung des akademischen Grades
Doctor rerum naturalium
(Dr. rer. nat.)
im Fach (Biologie)
eingereicht an der
Lebenswissenschaftlichen Fakultät
Humboldt Universität Berlin

von

Priyanka Banerjee

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekanin/Dekan der Lebenswissenschaftlichen Fakultät
Prof. Dr. Richard Lucius

Gutachter/innen:

1. Prof. Dr. Dr. h.c. Edda Klipp
2. Prof. Dr. Hermann-Georg Holzhütter
3. PD Dr. Robert Preissner

Tag der mündlichen Prüfung: 29.06.2016

Truth can be stated in a thousand different ways,

yet each one can be true
Swami Vivekananda

"The essence of life is gratitude." with this symbolic quote I would like to sincerely thank both my supervisors Prof. Dr. Hermann-Georg Holzhütter and PD Dr. Robert Preissner for their patience and supervision during my studies. With their valid skill set they have contributed to my development towards becoming an independent and critically thinking scientist. I have always been treated by them with an open-door policy and given time for scientific discussions. I need to extend my thanks to all the colleagues at AG Holzhütter, AG Preissner and members as well as PI's at the graduate school of 'Computational System Biology' (CSB). The reason I am in science is because of the passion of science infused in me by my school teacher Mrs Sipra Dutta and my ex-supervisor Prof. Rebecca Wade. I would like to thank both of them from my heart for giving me the best impression of the scientific world. My specials thanks to my colleagues Atolani, Gosia, Jevgeni, Mathias, Sophie and Vishal for providing a fun-filled, supportive and caring working environment. I would also thank my colleague Nikolaus Berndt for helpfully discussing his work on kinetic modeling and providing data used in the work which is presented in this thesis.

I also like to acknowledge all the contribution, supports and suggestions of all other great scientists, fellow PhD students and academics who worked as co-authors or lab mates. My special thanks to Vishal for helping me with proof reading and Sophie for the German translation of my summary.

I am indebted to the administrative support provided by Prof Dr. Edda Klipp and the scientific co-ordinators at CSB : Dr. Cordelia Arndt-Sullivan and Dr Uta Kaufhold.

I gratefully acknowledge the funding as scholarship received from The Deutsche Forschungsgemeinschaft (German Research Foundation) for my studies and scientific visits to different research labs in and outside Europe.

My heartiest thanks to my spiritual teacher revered Swami Atmasthananda Maharaj. I would like to offer my special thanks to Swami Baneshananda Maharaj and all members of the Vedanta society, for their constant spiritual and moral support.

Lastly, I would like to thank my family for all their love and encouragement. For my father, Mr Raghuji Banerjee and my mother, Mrs Alpnana Banerjee who raised me with a love for education and supported me in all my pursuits. My love and heartiest thanks to my elder sister Miss Mausumi Banerjee for being a terrific lady, supporting me in all my choices and always believing in me. I want to specially thanks my friends Ashish, Antonio, Anirban, Dmitry, Ejafa, Irina, Keerti, Leena, Marilena, Mahsa, Mansi, Malvika, Payal, Pratik, Roma, Ramesh, Sophie, Sahil, Tripti and Vihar for their time, love, support and for keeping me always motivated.

Contents

Contents iv

Zusammenfassung vii

Abstract ix

1 Introduction 1
- 1.1 Aim of the thesis . . . 3
- 1.2 Outline of the thesis . . . 4
- 1.3 Publications included in this thesis . . . 4

2 Theoretical background 7
- 2.1 *In silico* toxicology . . . 7
- 2.2 Cheminformatics . . . 8
  - 2.2.1 Molecular representations . . . 8
    - 2.2.1.1 Molecular graph . . . 9
    - 2.2.1.2 SMILES . . . 9
    - 2.2.1.3 SMARTS . . . 9
    - 2.2.1.4 InChI and InChIKey . . . 10
    - 2.2.1.5 Molecular fingerprints . . . 10
- 2.3 Molecular descriptors . . . 11
  - 2.3.1 Functional groups . . . 11
  - 2.3.2 Electronegativity and Partial charge . . . 11
  - 2.3.3 Octanol/water partition coefficient . . . 12
- 2.4 Pharmacophore . . . 12
- 2.5 Machine learning in computational toxicology . . . 13

3 Development of a novel method for prediction of rodent oral toxicity using similarity and fragment based approach 15
- 3.1 Introduction . . . 16
- 3.2 Data preparation and processing . . . 17
- 3.3 Molecular fingerprints . . . 18
  - 3.3.1 Path-based fingerprints (FP2 and FP4) . . . 18
  - 3.3.2 Circular fingerprints (ECFP) . . . 18
- 3.4 Method . . . 19
  - 3.4.1 Similarity method . . . 19
  - 3.4.2 Fragmentation method . . . 21

3.5 Results . . . 22
3.6 Performance evaluation . . . 22
3.7 Comparison with other methods . . . 24
3.8 Discussion . . . 26
3.9 Conclusion . . . 27
3.10 Application of the *in silico* prediction method in the development of natural product database. . . . 27
3.10.1 Toxicity prediction . . . 29
3.10.2 Graphical user interface and information retrieval . . . 31
3.11 Availability . . . 32

4 Development of binary classifiers for predictions of compounds active in toxicological pathways 34
4.1 Introduction . . . 35
4.2 Targets / Assays . . . 35
4.3 Data preparation and processing . . . 36
4.4 Molecular fingerprints . . . 37
4.4.1 Substructure fingerprints . . . 37
4.4.2 Circular fingerprints . . . 37
4.4.3 Estate fingerprints . . . 38
4.4.4 Toxicity fingerprints . . . 38
4.5 Molecular descriptors . . . 39
4.6 Methods . . . 40
4.6.1 Similarity-based fingerprint method . . . 40
4.6.2 Naive Bayes . . . 41
4.6.3 Random Forest . . . 42
4.6.4 Probabilistic Neural Network . . . 42
4.7 Constructions of the machine learning models . . . 43
4.8 Performance evaluation . . . 44
4.9 Results . . . 45
4.9.1 Results: Similarity-based fingerprint method . . . 45
4.9.2 Results: Naive Bayes . . . 46
4.9.3 Results: Random Forest . . . 47
4.9.4 Results: Probabilistic Neural Network . . . 48
4.10 Analysis of chemical space . . . 49
4.11 Comparison with Tox21 challenge top performers . . . 51
4.12 Discussion . . . 52
4.13 Conclusion . . . 54
4.14 Availability . . . 54

5 A novel method to predict fatty liver drugs using metabolic network based target identification 55
5.1 Introduction . . . 55
5.2 Metabolic Network . . . 56
5.3 Selection of target . . . 57
5.4 Fatty acid oxidation . . . 58
5.5 Formulation of hypothesis . . . 59
5.5.1 First hypothesis . . . 60

5.5.2 Second hypothesis . . . 60
5.6 Data set processing . . . 62
5.6.1 Pharmacophore model for drugs similar to malonyl CoA . . . 62
5.6.2 Computational docking of drugs similar to carnitine . . . 64
5.6.3 Known fatty liver drug set . . . 64
5.7 Methods . . . 64
5.7.1 Ligand based pharmacophore modeling of malonyl CoA similar inhibitors 65
5.7.2 Computational molecular docking based on carnitine similar inhibitors 68
5.8 Results . . . 69
5.8.1 Results of pharmacophore model . . . 70
5.8.2 Results of computational docking . . . 72
5.9 Predicted fatty liver drugs . . . 73
5.10 Discussion . . . 75
5.11 Conclusion . . . 76

6 Prediction of drugs interacting with HLA alleles 77
6.1 Introduction . . . 77
6.2 Methods . . . 79
6.2.1 Analysis of pharmacophoric groups . . . 79
6.2.2 Similarity based screening . . . 81
6.2.3 Molecular docking . . . 82
6.2.3.1 Analysis of docking results . . . 82
6.3 Results . . . 83
6.3.1 Computational validation . . . 84
6.3.2 Experimental validation . . . 86
6.4 Conclusion . . . 86

7 Summary 88
7.1 Summary . . . 88
7.2 Discussion and conclusion . . . 89
7.3 Limitations of the methods presented in this thesis . . . 94
7.4 Perspectives and future work . . . 95

List of Figures 95

List of Tables 99

Abbreviations 101

A Software and databases 118

B Ehrenwortliche Erklarung 120

Zusammenfassung

Die Menschheit ist vielfältigen chemischen Wirkstoffen ausgesetzt – zum Beispiel durch Kosmetika und Pharmazeutika sowie durch viele andere chemische Quellen. Es wird angenommen, dass diese stetige Exposition mit Chemikalien gesundheitliche Beeinträchtigungen bei Menschen hervorruft. Zudem haben Regulierungsbehörden aus Europa und den USA festgestellt, dass es ein Risiko gibt, welches mit der kombinierten Exposition durch mehrere Chemikalien im Zusammenhang steht. Mögliche Kombinationen von Tausenden Wirkstoffen zu testen, ist sehr zeitaufwendig und nicht praktikabel. Das Hauptanliegen dieser Arbeit ist es, die Probleme von Off-target-Effekten chemischer Strukturen zu benennen – mit den Mitteln der Chemieinformatik, der strukturellen Bioinformatik sowie unter Berücksichtigung von computerbasierten, systembiologischen Ansätzen.

Diese Dissertation ist in vier Hauptprojekte eingeteilt, die jeweils für vier unterschiedliche computerbasierte Simulationsmethoden stehen, um verschiedene Ebenen toxikologischer Untersuchungen zu adressieren.

Im Projekt I (Kapitel 3) wurde ein neuartiger Ensemble-Ansatz basierend auf der strukturellen Ähnlichkeit von chemischen Wirkstoffen und Bestimmungen von toxischen Fragmenten implementiert, um die orale Toxizität bei Nagetieren vorherzusagen. Dieser Ansatz brachte überzeugende und konsistente Ergebnisse, die bereits vorhandenen kommerziellen sowie anderen bekannten Methoden überlegen sind. Im Projekt II (Kapitel 4) wurden – auf der Grundlage von Daten des Tox21 Wettbewerbs – unterschiedliche Machine-Learning Modelle entwickelt und verglichen, um die Komponenten vorherzusagen, die in den toxikologischen Stoffwechselwegen mit Zielmolekülen interagieren. Die vorgestellten Methoden dieser Studie können von Regulierungsbehörden genutzt werden, um im großen Umfang die Toxizität von target-spezifischen Wirkstoffen vorherzusagen.

In Projekt III (Kapitel 5) wird ein neuartiger Ansatz beschrieben, welcher das dreigliedrige Konzept aus computerbasierter Systembiologie, Chemieinformatik und der strukturellen-Bioinformatik nutzt, um Medikamente zu bestimmen, welche das metabolische Syndrom hervorrufen. Zwei unterschiedliche, neuartige Mechanismen für das medikamenteninduzierte Fettleber-Syndrom wurden identifiziert und computerbasiert validiert.

In Projekt IV (Kapitel 6) wurde in silico ein Screening Protokoll entwickelt, welches die strukturelle Ähnlichkeit, die pharmakophorischen Eigenschaften und die Überprüfung von computerbasierten Docking Studien berücksichtigt. Dieser Ansatz führte zur Identifikation neuer Bindungsstellen für Acyclovir in der Peptid-Bindungsstelle für das HLA-spezifische Allel. Solche spezifischen Verbindungen von Medikamenten und dem HLA-spezifischen Allel verursachen immun-modulierte unerwünschte Medikamentenwechselwirkungen.

Der Gesamtbeitrag dieser Arbeit beinhaltet die Entwicklung unterschiedlicher computerbasierter Methoden mit speziellen chemieinformatischen Ansätzen, welche erfolgreich eingesetzt werden können, um die Probleme von Off-target Effekten auf unterschiedlichen Ebenen des Systems zu verstehen.

Schlagwörter: Toxizität Vorhersage, Maschinelles lernen, Cheminformatik, toxischen Fragmenten, computerbasierter Systembiologie,

Abstract

Exposure to various chemicals agents through cosmetics, medications, preserved food, environments and many other sources have resulted in serious health issues in humans. Additionally, regulatory authorities from Europe and United States of America have recognized the risk associated with combined exposure to multiple chemicals. Testing all possible combinations of these thousands of compounds is impractical and time consuming. The main aim of the thesis is to address the problem of off-targets effects of chemical structures by applying and developing cheminformatics, structural bioinformatics and computational systems biology approaches.

This dissertation is divided into four main projects representing four different computational methods to aid different level of toxicological investigations.

In project I (chapter 3) a novel ensemble approach based on the structural similarity of the chemical compounds and identifications of toxic fragments was implemented to predict rodent oral toxicity. This approach showed powerful and consistently best performance over available commercial as well as other public methods.

In project II (chapter 4) different machine learning models were developed and compared using Tox 21 challenge 2014 data, to predict the outcomes of the compounds that have the potential to interact with the targets active in toxicological pathways. The methods proposed in this study can be used by the regulatory agencies to access the toxicity of these target specific compounds in large scale.

In project III (chapter 5) a novel approach integrating the trio concept of 'computational system biology, cheminformatics and structural bioinformatics' to predict drugs induced metabolic syndrome have been described. Two different novel mechanisms for drug induced fatty liver syndrome has been proposed and computationally validated.

In project IV (chapter 6) a *in silico* screening protocol was established taking into the structurally similarity, pharmacophoric features and validation using computational docking studies. This approach led to the identification of novel binding site for acyclovir in the peptide

binding groove of the human leukocyte antigen (HLA) specific allele. Such specific binding of drugs to the specific HLA alleles results in immune-mediated adverse drug reactions.

The over all contribution of the thesis includes development of different computational methods that can successfully be applied to address the problems of off-targets effects at various levels of system inference.

Keywords: Toxicity prediction, machine learning, fatty liver, system biology, similarity searching, cheminformatics

Chapter 1

Introduction

"Poison is in everything, and no thing is without poison. The dosage makes it either a poison or a remedy". Almost 500 years ago, Paracelsus acknowledged that the subtle distinction between whether a given chemical compound acts as a medicine or poison is often determined by the dose as well as amount of time at which it is given[1]. However, it is only by the twentieth century Paracelsus understanding of dose-dependent toxicity profiles of chemical compounds manifested a new definition in terms of system level understanding of toxicity in modern science[2]. From the science of poisons and intoxication, toxicology and its scope has broadened and developed with time, and currently has been regarded as the 'science of safety' and 'personalized therapeutic administration'[2]. In summary, toxicology is a translational science, transferring knowledge from basic science into practical applications to safe guard public health and the environment[2].

Humans are exposed to various chemicals agents through cosmetics, pharmaceuticals and many other sources. Exposure to chemicals is suspected as playing a main role in development of some adverse health effects in humans. Mechanism-based prediction and evaluation of chemical compound's toxicity constitute an evolving science whose development is critical to drug discovery, development and regulatory evaluation[3]. The vision involved in this outlook is to make advances in human therapeutics available while protecting public health.

There are different ways a toxic response is associated with chemical compound (e.g drugs, environmental chemicals). Toxic effects are sometimes direct and predictable following an overdose of the chemical compounds. This effect can be due to direct association between presence of certain toxic patterns (fragments) in the chemical compounds and toxicity[4]; on the other hand it may be sometimes due to changes in the metabolism of the compound resulting in toxic metabolites[5]. Additionally this direct and predictable toxic response may be mediated by metabolites and may be pharmacolgic or immunologic in nature. It is also

observed that sometimes toxic response can be direct and yet unpredictable either following one dose or just a few doses. These effects are often termed as 'the idiosyncratic response'. They may appear only in a specific group of people and often there is no good prediction for their occurrence[1].

Chemical compounds and toxic effects can be well spread including hierarchically levels connecting one functional domain to the next, starting with initial interactions with molecular targets to organelles (such as mitochondria) to cells and to organs, resulting in detectable as well as measurable toxic response[3]. Often chemical compounds (drugs) along with their desired therapeutic effects can also result in off-target effects and systemic adverse reactions, due to several interactions with other molecular targets in the target network.

Rapidly evolving data resources from various levels of information such as chemistry, structural bioinformatics, cheminformatics, metabolomics, proteomics, genomics and their interconnections has been the foundation of the predictive science. The growth in the field of *in silico* toxicity prediction also known as computational toxicology has led to the development of methods and models for prediction of toxicological endpoints, clinical adverse effects and metabolism of pharmaceutical substances[6]. Additionally regulatory authorities from Europe and the United States of America have realized the risk associated with combined exposure to multiple chemicals. Testing all possible combinations of these thousands of compounds is impractical and time consuming. Thus computational toxicity prediction methods integrate all relevant information on a compound and its structural or substructural analogs to make preliminary assessments of potential toxicity. Furthermore, along with the identification of the chemical determinants that are associated with observed off-target effect, computational methods can also help to investigate the underlined mode of action or molecular mechanism[7] .

This thesis focuses on the development of *in silico* methods for prediction of toxic outcomes in chemical compounds. The thesis envisages a global view in response to toxicity profiles of chemical compounds as well as local (specific) consideration on the mechanism and pathway level. For example, a drug or a chemical compound might interact with a molecular target which can result in interactions with multiple molecular targets including both therapeutic as well as off-target with different affinities. In this process, consequently it can activate different signaling pathways or interact with functional pathways. Additionally, such interactions at cellular level can produce toxic effects on certain organs. This can be further extended to the adverse drug reactions (ADRs) profile of population sharing similar toxicological pathways or network. A schematic diagram of such view is represented in the figure 1.1 .

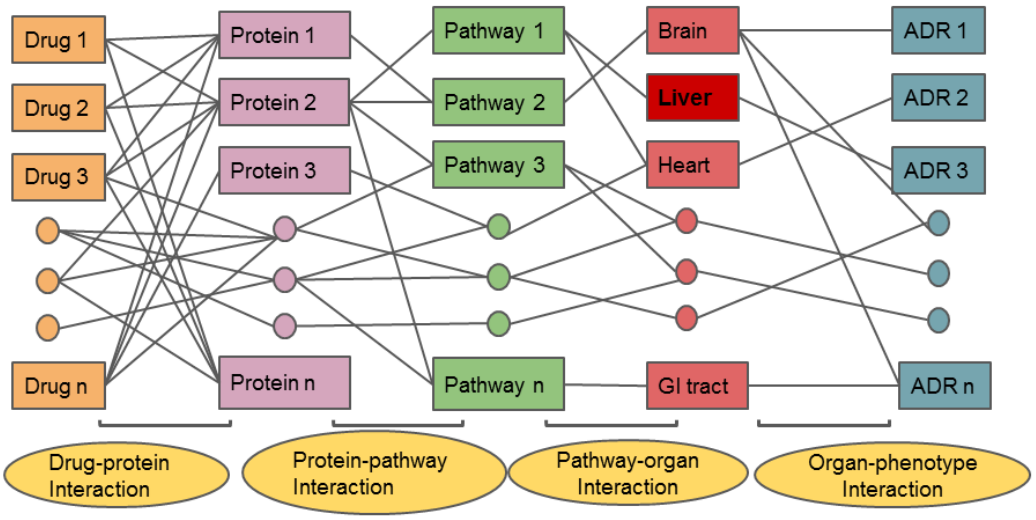


FIGURE 1.1: **Different layers of information considered in this thesis to address off-targets effects** The figure is inspired from[6] and represents drugs interacting with different targets, can effects several pathways resulting in off-targets effects on specific organs, sometimes leading to specific ADRs profile.

1.1 Aim of the thesis

The dissertation study represented here tries to address the following major questions :

1) Can we predict compound toxicity based on rules that define the relationship of structure and toxicity from existing data ?

2) Can we derive quantitative correlations between compound structure and off-target effect using machine learning methods and rationally selected molecular descriptors, interacting with targets involved in toxicological pathways ?

3) Can we integrate the concepts of kinetic modeling of metabolic network, structural bioinformatics and cheminformatics to predict drugs with fatty liver profiles and the underlying mechanism of drug induced fatty liver ?

4) Can drugs having similar structural and therapeutic classes bind to specific HLA allele in similar manner ?

1.2 Outline of the thesis

The thesis structure is organized as follows:

Chapter 1, introduces the reader to the subject and aims of this PhD dissertation.

Chapter 2 provides the important 'theoretical background' of some key concepts necessary to understand the discussed topics and results in this thesis.

Chapter 3 describes a novel computational method for the prediction of rodent acute oral toxicity. The uniqueness of this approach includes identification of certain toxic patterns along with similarity approach to predict toxic outcomes only based on chemical structures. Additionally a natural product database is reported along with toxicity profiles as well as pathways associated with the natural compounds.

Chapter 4 describes the development of different computational models to identify chemicals that could be active in toxicological pathways.

Chapter 5 represents a novel integration of the trio 'computational systems biology , structural bioinformatics and cheminformatics' approaches to predict drugs involved in metabolic syndrome and the possible underlying mechanism.

Chapter 6 describes a virtual screening protocol and computational method to identify approved drugs that can bind to HLA specific alleles in the context of collaboration resulting in the *in silico* identification of experimentally verified approved drug that binds to HLA specific alleles.

Chapter 7, summarizes the main results and important finding in this thesis. Concluding remarks for the impact of this dissertation are also highlighted along with possible areas for future studies.

1.3 Publications included in this thesis

This thesis consists of the following publications. The author's contribution to each of the publication is stated below.

Chapter 3 is based on

1. Drwal M, **Banerjee P**, Dunkel M, Wettig M and Preissner R

   'ProTox: A web server for the in silico prediction of rodent oral toxicity' Nucleic Acids Res. 42(Webserver issue W1) : W53-58.

(The author was involved in design of the project. The author contributed equally in data preparation and analysis, similarity method development and calculations of the descriptions. Fragmentation method was implemented by the author. Additionally, author was involved in the design of the web-server)

2. **Banerjee P**, Drwal M and Preissner R

   Identification of chemical fragments involved in toxicity (in prep)

   (The author has designed and conceived the study. Performed all the analysis, developed the method and written the manuscript)

3. **Banerjee P**, Erehman J, Gohlke B.O, Wilhelm T, Preissner R and Dunkel M

   Super Natural II: A database of natural products Nucleic Acids Res.(Database issue) (2014)

   (The author was involved in design of the study. Generation of data and analysis. Included toxicity prediction part to the database. Involved in the development of the database and written the manuscript).

Chapter 4 is based on

1. Drwal MN, Siramshetty VB, **Banerjee P**, Goede A, Preissner R and Dunkel M (2015)

   Molecular similarity-based predictions of the Tox21 screening outcome Front. Environ. Sci. 3:54

   (The author introduced the concept of KNIME based modeling in this project and initiated the machine learning model approach. The author was equally involved in the data preparation and analysis, generation and validation of the predictive models, calculations and selections of the descriptors.)

2. **Banerjee P**, Siramshetty VB, Drwal M and Preissner R (2016)

   Computational methods for prediction of in vitro effects of new chemical structures using Tox21 challenge data J Cheminform 8:51.

   (The author equally conceived the study. Designed, developed and analyzed the machine learning models and chemical space analysis. Author was involved equally in the manuscript writing process along with Siramshetty VB)

Chapter 6 is based on

1. Metushi I.G, Wriston A, **Banerjee P**, Gohlke B.O, English A.M, Lucas A, Moore C, Sidney J, Buus S, Ostrov D.A, Mallal S, Phillips E, Shabanowitz J, Hunt D.F, Preissner R and Peters B

Acyclovir Has Low but Detectable Influence on HLA-B*57:01 Specificity without Inducing Hypersensitivity. PLoS ONE 10(5): e0124878 (2015).

(The author was involved with computational part of the study. The author along with B.O Gohlke and R. Preissner conceived the study. The author analysed the pharmocophoric groups, equally involved in 2D similarity search strategy, performed the molecular docking studies and analysis and written the computational part along with pictures generation).

Chapter 2

Theoretical background

2.1 *In silico* toxicology

In silico toxicology, also referred as computational toxicology is a novel strategy in the field of toxicological research which aims to establish mathematical models based on existing knowledge for predictions of untested and possibly unsynthesised, compounds[2]. One of the most important features of computational toxicology is its high degree of integration of different layers of informations linking computational systems biology, bioinformatics and cheminformatics methodologies. Given an ideal case, it is possible to establish the toxicity profiles of chemical structures associated with known mechanism of action (MOA) or even better link with an 'toxicological pathways'. However, if the toxic effect is the result of several different mechanism working sequentially or simultaneously, then the reliable prediction based on the chemical structure could be difficult[3] .

Computational toxicology is highly dependent on the quality as well as large amount of data. Therefore, data extracted from standard experimental conditions can significantly reduce the error associated with the predictions of the *in silico* models. Although computational toxicology have developed recently, however the increase in publications in literature databases (such as PubMed) over the last decade on topic as '*in silico*' and 'toxicity' highlights their growing importance. The funding for projects related to toxicity based research through EU grants has been rising at alarming rate such as EU-ToxRisk (http://www.eu-toxrisk.eu/, COSMOS - a European Union project developing methods for determining the safety of cosmetic ingredients for humans, without the use of animals, using computational models (http://www.cosmostox.eu/), and NOTOX (http://notox-sb.eu/). On the other hand, TOXCAST (http://www.epa.gov/chemical-research/toxicity-forecasting and Tox21 (http://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21)

platforms promoted by the US Environmental Protection Agency (EPA) in support of *in silico* toxicology contributed to the paradigm change away from animal testing to alternative methods by combining system biology, molecular biology and computational methods.

2.2 Cheminformatics

The analysis and methods developed in this thesis are based on the scientific discipline cheminformatics or Chemoinformatics, also sometime refers as chemical informatics or molecular informatics. Though there have been some disagreement over the name within the community, however 'cheminformatics' is favored by the Journal of Cheminformatics (http://jcheminf.springeropen.com/) . Chemoinformatics is an emerging science that concerns the computational storage, retrieval and reasoning about chemical information[8] . Cheminformatics was first formally defined as the use of information technology and management has become a critical part of the drug discovery process. Cheminformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better and faster decisions in the area of drug lead identification and organization as stated by Brown in 1998[9]. This definition as termed by an industrial pharmaceutical scientist was obviously based towards the pharmaceutical applications. However with the advancement of research in the areas of chemistry, computational chemistry, molecular modeling, computer science and statistics, supports that recent cheminformatics based scientific research appeals to almost any area of chemical research, biological research, toxicological research and systems biology[10]. Cheminformatics addresses a broad range of problems in chemistry and biology; however, the most commonly known applications of chemoinformatics approaches have arguably been in the area of drug discovery, where cheminformatics based methods have tried to establish robust relationships between a chemical structure and its physical or biological properties[8] .

2.2.1 Molecular representations

Large amount of chemical as well as biological data are stored in databases. Some of these databases are publicly available while some are commercially[11] . They contains large number of compounds ranging from several hundred thousands to million of entries. It is nevertheless possible to search data in these databases and retrieve the information in seconds. In this section, some of the most commonly used methods in cheminformatics to represent molecular structures and used in the creation of structural databases are described.

2.2.1.1 Molecular graph

Molecular graph is an usual way to store chemical structures in a computer. Graph theory has been well-established in the area of mathematics which has found its recognition not only in chemical or biological sciences but in also many other areas, such as computer science[13]. Chemical structures can be effectively represented as labeled graph[12]. A graph consist of *nodes* and *edges*. In a molecular graph each atom is associated with a node and each bond with an edge. The nodes and edges in a molecular graph may have properties associated with them. That is, the nodes may represent the atomic number, symbol where as the edge represents the bond order. These properties are important while performing some operations using the molecular graph. As these molecular graphs can be constructed in more than one ways, it is therefore important to have methods that can determine whether two molecular graph are identical. Theoretically, this problem is known as *graph isomorphism*[13] .

2.2.1.2 SMILES

SMILES (Simplified Molecular Input Line System) is an unambiguous and reproducible method for represent molecules computationally developed by Wiswesser in the year 1985[14]. This method represent molecule structure as a linear string of symbols based on certain predefined rules. SMILES are based on hydrogen suppressed graph based on the assumption that hydrogens contributes to atom's lowest normal valence. All the non-hydrogen atom in a SMILES string are represented by their atomic symbols enclosed in square brackets. Formal charges are assigned as + or - ; where as aromatic atoms are specified using the lowercase atomic symbols. Single, double , triple and aromatics bonds are represented as '-', '=', ' #' and ':,' respectively. Cyclic structures are represented by breaking a ring at a single or aromatic bond and numbering the atoms on either side of the break with a number. For example, benzene (cyclohexane) can be represented as SMILES string C1CCCCC1[15].

SMILES are efficient ways for storage and retrieval compounds across multiple computer platforms. SMILES algorithm can detect most of the aromatics compounds with applied Huckel's rule, but do not account for the tautomers in the chemical structures. Therefore, there can be more than one valid SMILES for chemical structures based on the algorithm used to create them.

2.2.1.3 SMARTS

SMARTS (SMILES Arbitrary Target Specification) is an extension of SMILES that allows for variability within the represented chemical structure. It encodes based on all the rules of SMILES and additionally includes logical operators. such as 'AND' (&), 'OR' (,), and 'NOT'

(), and special atomic and bond symbols that provide flexibility to chemical names. SMARTS representation of molecules are often used for substructure searching[16] .

2.2.1.4 InChI and InChIKey

InChI (International Chemical Identifier) was developed and released in 2005 as open source chemical structure representation algorithm with an aim to established a common platform to unify searches across multiple chemical databases. It is maintained by the InChI Trust (http://www.inchi-trust.org/. InChIKey is the hashed key version of InChI. The aim of the InChIKey is to provide a unique representation of chemical structures that can be indexed for searching in databases[17].

InChI is made up of several layers to represent different classes of information in a chemical structure. The starting first two layers encodes general information like chemical formula and connections between atoms[18]. Additional layers encodes the information on stereo-chemistry, tautomerism and isotopic information. The three sublayers represents the all bond to non-bridging hydrogen atoms„ immobile hydrogen atoms and mobile hydrogen atoms respectively[18]. The algorithm behind InChI creation includes six normalization rules taking into consideration of variable protonation and identification of tautomeric patterns and resonances to establish an unique and consistent chemical structure representation[18].

InChIKey generates two blocks using a truncated SHA-256 cryptographic hash function based on InChI. The keys contain a fixed length of 27 characters making sure there is a minimal chance of two different molecules with the same hash key. Use of InChIKeys for comparing molecules in different data bases were tested and obtained very low error rate[19].

2.2.1.5 Molecular fingerprints

Molecular fingerprints encodes the structural, pharmacophore or property descriptors in to binary bit string format[20] . Each bit position accounts for the presence or absence of a given pattern. The bit is set to '1', if a particular pattern is present in the molecule and set to '0' if the pattern is absent[21]. There are different molecular fingerprints types varying substantially in their design and length. Mostly these fingerprints have fixed length, but not always. In addition to the presence and absence of a given pattern, fingerprints also account for the number of occurrences of the pattern in the molecule (e,g hydroxyl group or a benzene ring). Molecular fingerprints representations are often used to search for molecules 'similar' to a query compound[22].

2.3 Molecular descriptors

Molecular descriptors can represent both structural or physicochemical properties. Based on 1D; 2D and 3D co-ordinates of a chemical structures descriptors can encodes properties like molecular weight, geometry, chemical formula, volume, surface areas, ring content, rotatable bonds, interatomic distances, bonds, atom types, electronegativities, polarizabilities, symmetry, atom distribution, topological charge, functional group compositions, aromaticity. solvation properties and many more[23]. All these descriptors are computed using knowledge-based, graph-theoretical methods, molecular mechanical or quantum-mechanical tools[23].

Higher information content encoded by these descriptors makes them extremely important important and useful for model generation[24].

Some of the important descriptors used in this study are explained in details.

2.3.1 Functional groups

According to the International Union of Pure and Applied Chemistry, functional groups are atoms or groups of atoms that have similar chemical properties across the range of various chemical compounds[25]. They are attached to a central backbone often called as scaffold or chemotype defining its physical and chemical properties. As a consequence, the location and nature of functional groups for a compound contains key information related to its activity[25]. There are different functional groups such as hydrocarbons, halogens, oxygen, nitrogen, sulfur, phosphorous, amides, alcohols, esters, ethers, nitro group, thiols etc[25].

Functional groups can either can be represented based on their atomic composition and bonds or implicitly encoded by their general properties. To explain, it is often observed that under physiological conditions carboxyl groups are often negatively charged. This property is reflected in the structure of the functional group and this property corresponds to the most important information on the biochemical activity of a given chemical compound[25]. On the other hand, the capacity to form hydrogen bonds can greatly influence a molecule's properties. Hydrogen bonding interactions influence the electron distribution of neighboring atoms and the site's reactivity; this makes it an important functional property for protein-ligand interaction[26].

2.3.2 Electronegativity and Partial charge

This descriptor encodes the charge distribution over an entire molecule. To assign partial charges to individuals atoms a new method was developed by Gasteiger and Marsili (1980) called a Partial Equalization of Orbital Electronegativity (PEOE)[27] .

2.3.3 Octanol/water partition coefficient

LogP (logarithm of partition coefficient between n-octanol and water) is regarded as one of the important molecular descriptors that has been successfully used in the field of drug design, virtual screening and several activities prediction methods[28].

2.4 Pharmacophore

A pharmacophore summarizes the steric and electronic features that are needed to ensure the optimal supramolecular interactions with a specific molecular target structure to trigger (or to block) its desired biological response[29]. Pharmacophore is a purely abstract representation of molecules that accounts for common molecular interactions capabilities of a group of compounds towards a given target. Some of the pharmocophoric descriptors used to define a pharmacophore are hydrogen bond donor, hydrogen bond acceptor, aromatic ring, hydrophobic and many more[29].

The standard eleven pharmacophoric features available at Discovery Studio are listed below:

1. **Hydrophobic** features represents a groups of atoms that are not adjacent to any charged atoms or electronegative atoms, but have surface accessibility including phenyl, cycloalkyl, isopropyl and methyl.
2. **Hydrophobic (aliphatic)** features represents only aliphatic atoms.
3. **Hydrophobic (aromatic)** features represents only aromatic atoms.
4. **Hydrogen bond acceptor** features represents the sp or sp3 nitrogen and sp3 oxygen or sulphur that have a lone pair and charge less than or equal zero. Additionally basic amines that have a lone pair.
5. **Hydrogen bond acceptor (lipid)** features represents atom or groups of atoms like nitrogen, oxygen or sulphur that have a lone pair and charge less than or equal to zero.
6. **Hydrogen bond donor** features represents only atom or groups of atoms like non-acidic hydroxyls, acetylenic hydrogens, thiols, NHs (except tetrazoles and trifluromethyl sulfonamide hydrogens).
7. **Negative charge** features represents only atom or groups of atoms that have negative charges not adjacent to a positive charge.

8. **Negative ionizable** features represents atoms or groups of atoms that are likely be deprotonated at physiological pH, such as trifluoromethyl sulfonamide hydrogens, phosphonic acids, sulfinic acids, phosphinic acids or caboxylic acids, tetrazoles, sulfonic acids.

9. **Positive charge** features represents atom or groups of atoms that have positive charge not adjacent to a negative charge.

10. **Positive ionizable** features represents atoms or groups of atoms that are likely to protonated at physiological pH, such as basic amines, basic primary amidines, basic secondary amidines, basic guanidines.

11. **Ring aromatic** features represents aromatic rings with five or six member atoms.

Ligand based pharmacophore model is based on the HipHop algorithm. The algorithm considers the 3D spatial arrangements of the chemical features that are common between the set of active ligands for a chosen target[29]. The features are identified by a pruned exhaustive search initiating with small sets of features and extending to large number of common features. The molecules in the training set are evaluated on the basis of the types of chemical features they contain, along with their ability to adopt a confirmation that allows the chosen features to be superimposed on an optimal configuration[29]. The number molecule that need to be mapped with the common features in the training set is user defined. The resultant pharmacophores are ranked based on the fit value as well as the maximum number of the training set that matched with the given features. The pharmacophore model can be validated with a external validation set which can contains both actives and in-actives and their specificity and sensitivity can be evaluated[29].

2.5 Machine learning in computational toxicology

Machine Learning can be defined as a collection of computational approaches used in predictive science given a tagged datasets[30]. Machine algorithms are generally developed in computer science and other related disciplines and have found association with chemical modeling by a process of diffusion. In general, datasets of molecular structures are tagged based on their activity; like active and inactive and this information is utilized to model a binary classification based on the selection of appropriate molecular descriptors[30]. The selection of the molecular descriptors can be categorized into two steps: firstly, the molecular structures are classically represented in form of a molecular graph and is converted into feature vectors into the chemical space. Secondly, the features vectors are mapped with the property of interest using a mathematical function[30]. Usually, the mapping of these descriptors is often learned by the machine learning algorithm which is later used to make decisions.

Machine learning methods in the field of the cheminformatics have been widely used in the areas of bioactivity and ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties prediction[31,32]. It has been demonstrated that models build on machine learning methods which take into account large dimensional data are highly successful for robust external predictions. Machine learning and pattern recognition has a long association with chemistry counting back for more than four decades, such as in the field of spectroscopic data interpretation[30].

Although there are multiple machine learning algorithms which has been used in the field of predictive modeling, nevertheless in this thesis three most popular classification algorithm were implemented ; Naive Bayes (NB), Random Forest (RF) and Probabilistic Neural Network (PNN). All the three algorithms are explained in details in the chapter 4 .

Chapter 3

Development of a novel method for prediction of rodent oral toxicity using similarity and fragment based approach

The drug discovery and development process implies the design and selection of compounds for identified therapeutic molecular target (proteins) and optimization of these compounds to enhance activity. However, many compounds synthesized during the lead optimization phase have failed to successfully place into the market mainly due to the discovery of adverse effects[7]. Therefore, a critical priority in the process of drug development and safety assessment conducted by the regulatory authorities includes the early screening and detection of serious toxicological issues[7]. Although, *in vivo* toxicology has been applied intensively for identification of side effects induced by a drug, however it is found that this approach alone cannot help to reduce the large failure rate in late-stage clinical trials[7]. It is indeed important to have methods that can be easily integrated into the early stage of the drug discovery process using only 'chemical structure' of the compounds. Regulatory agencies like the U.S Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have taken initiative to improve the prediction of toxicity of new molecules in the drug discovery and development pipelines[33]. The Replace, Reduce, Refines (3R) principle (European partnership for alternative approaches to animal testing) has been able to encourage the development and optimization of alternative methods including *in vitro* assays and in silico predictions in recent years[7].

In general, *in silico* toxicology can be defined as mathematical data analysis and development of computational algorithms for the prediction of toxicological activity of a compound.

A large number of *in silico* models have been developed which focus on different levels of toxicity such as on system level, specific organs, biochemical mechanism or certain biological processes[34]. The U.S Environmental Protection Agency (EPA) has defined computational toxicology (i,e in silico toxicology) as the''integration of modern computing and information technology with molecular biology to improve agency prioritization of data requirements and risk assessment of chemicals"[35].

This chapter describes the similarity and fragment based ensemble method developed to predict the lethal dose (LD_{50}) values for oral toxicity in rodents using only the chemical space[36].

3.1 Introduction

Animal trials have been used for over 60 years to evaluate toxic profiles of chemicals used as medicines, cosmetics, food additives, industrials and agricultural chemicals[36]. Use of animals in toxicity issues has been expensive, time consuming and associated with ethical concerns. Thus, *in silico* prediction methods have found as an alternative approach and aim to rationalize the preclinical drug development reducing the time and costs involved in animal experiments. The prediction method described in this section is based on the structural similarity of compounds with known median lethal doses (LD_{50}) values and incorporation of identified fragments (structural alert) over-represented in toxic compounds, representing a novel approach in toxicity prediction.This method was validated on an external evaluation set and displayed a strong performance in terms of sensitivity, specificity and precision of 76 %, 95 % and 75 % respectively. There are several commercial as well as freely available methods available for *in silico* toxicity prediction[36]. Some of the most used commercial methods (softwares) in this domain are Discovery Studio's TOPKAT (Toxicity Prediction by Komputer Assisted technology; Accelrys, Inc., USA) (http://accelrys.com/) , ADMET Predictor (Simulations Plus, Inc., USA) (http://www.simulations-plus.com/) and ADME-Tox Prediction (Advanced Chemistry Development, Inc., Canada) (http://www.acdlabs.com/). Publicly available methods includes the Toxicity Estimation Software Tools (T.E.S.T) (http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test) developed by the U.S Environmental Protection Agency and AdmetSAR (web server), and are mainly based on QSAR method. In spite of achieving good rates in toxicity prediction, QSAR method have certain disadvantage for prediction of structurally novel compounds which was not previously included in the training set[36]. Moreover, in order to include new data into the training set, redevelopment of new QSAR model is necessary. On the other hand implementation of similarity based methods are computationally inexpensive and new compound data can be easily added to the training set[36]. The essence of the similarity based method has been

the assumption that molecules that are globally similar in structure should exhibit similar biological activity[37]. That is molecules exhibit 'neighborhood behavior'[38]. Similarity methods have been successfully applied in the areas of compound selection for drug discovery, drug repositioning studies and the prediction of drug-target interactions[39] . Thus, an approach was designed based on integration of compound similarity (global) and fragment (local) occurrences for toxicity prediction.

3.2 Data preparation and processing

The data set used in this study was obtained from an in-house database of toxic compounds 'SuperToxic'[40]. The data set used for the method development consists of approximately 38 000 unique compounds with known oral LD_{50} values in mg/kg body weight measured in rodents. The complete dataset was standardized using Instant JChem 6.2.0 (January 2014), ChemAxon. InCHI keys were calculated using the Discovery Studio version 4.2 and used to remove duplicates in the dataset. In case a compound was reported with multiple LD_{50} values, the lowest value was considered to represent the worst-case toxicity. Six toxicity classes are defined according to the globally harmonized system (GHS) (https://www.osha.gov/dsg/hazcom/global.html) of classification of labeling of chemicals as shown in the table below:

Toxicity class	**Definition and LD_{50} in mg/kg**
I	fatal if swallowed ($LD_{50} \leq 5$)
II	fatal if swallowed ($5 < LD_{50} \leq 50$)
III	toxic if swallowed ($50 < LD_{50} \leq 300$)
IV	harmful if swallowed ($300 < LD_{50} \leq 2000$)
V	may be harmful if swallowed ($2000 < LD_{50} \leq 5000$)
VI	non-toxic ($LD_{50} > 5000$)

TABLE 3.1: **Toxicity classes with definition and LD_{50} value range**.

The distribution of training set molecules in different toxicity classes is provided below:

Toxicity class	**number of molecules**
I	508
II	1886
III	6808
IV	20987
V	6343
VI	1985

TABLE 3.2: **Number of compounds per toxicity class in the training set** .

3.3 Molecular fingerprints

Molecules were represented as molecular fingerprints for similarity calculation. For general description of molecular fingerprints see in section 2.2.1.5 . Two different fingerprints : concatenated 'FP2' and 'FP4' fingerprints and Extended Connectivity Fingerprint (ECFP)[41] are used in this study .

3.3.1 Path-based fingerprints (FP2 and FP4)

In a path-based fingerprint 'FP2', small molecules linear fragments are indexed in to a binary string format. These linear fragments length ranges from 1 to 7 atoms. However, single atom fragments (e,g Carbon (C), Nitrogen (N), and Oxygen(O)) are ignored[36]. In this kind of fingerprints, the ring structures are not encoded as ring, but stored as single canonical linear fragment. Each fragment is assigned as hash number from 0 to 1020 which is used to set a bit in a 1024 vector. Where as in case of 'FP4' fingerprints are created from a set of SMARTS patterns which takes into account of functional groups. These two fingerprints were calculated separately using Open Babel library. Later, both the fingerprints were concatenated into a single fingerprint which encodes the linear fragments as well as the functional groups for each of the molecule in the dataset[36].

3.3.2 Circular fingerprints (ECFP)

Circular fingerprints are the representation of molecular structures by means of circular neighborhoods. These fingerprints are a refinement of the Morgan algorithm[42], designed to identify the presence of particular substructures in a molecule. ECFP is generated by systematically recording the neighborhood of each non- hydrogen atom into multiple circular layers up to a given user defined diameter; by using a fixed hash function[41]. The results obtained from these hash functions are mapped into integer indices. Each index of the fingerprint denotes the presence of a particular substructure. The substructures size represented by each index depends on the number of layers used in the fingerprint generation process[41]. The layers are often termed as 'radius'. The substructures encoded in this fingerprints are not predefined and generated in a molecule dependent manner and therefore, represents huge number of different molecular features such as stereochemical information[41]. They have often been used in the field of toxicity prediction and believed to provide more adequate results for similarity searching[43]. The 'ECFP' fingerprint for this study was generated using the Java library provided by JChem 6.1.3 (November 2013), ChemAxon. A separate Java program was developed and the diameter for the creation of ECFP was considered as 4.

3.4 Method

The prediction method described in this study is a combination of two approaches; similarity and fragment-based approach. Each method is discussed in individual section below.

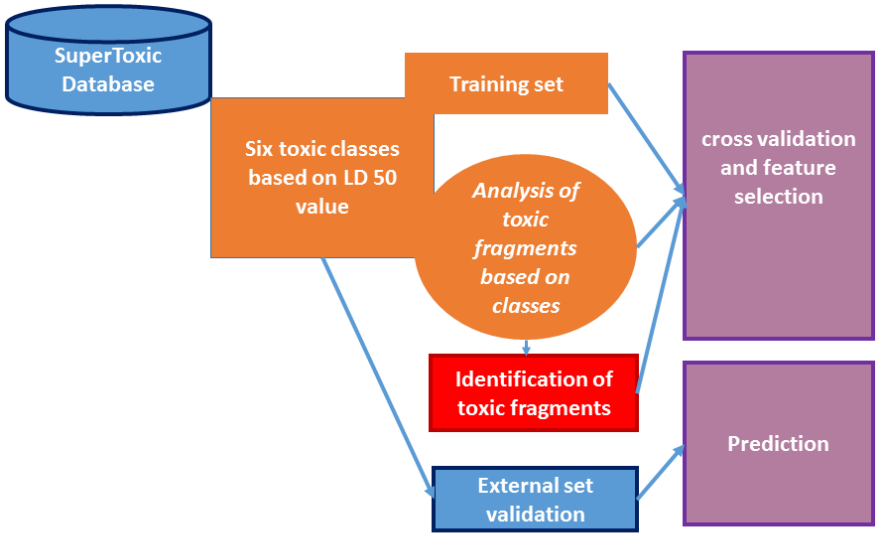


FIGURE 3.1: **Ensemble method** The workflow represents the data source as well as toxicity classes used in this study. The integration of the global similarity scores and local fragment based approach used for the toxicity prediction of the chemical compounds.

3.4.1 Similarity method

Similarity search calculation is performed based on the assumption of *'similarity property principle'* (i,e structurally similar molecules should have similar biological activity). Since, interaction of a small molecule and a target protein is based on their structures, so small molecules with similar structures are believed to interact in a similar manner with the protein target. The similarity score is computed comparing the two-dimensional (2D) molecular fingerprints of the molecules as described in the sections 3.3.1 and 3.3.2. The similarity score between two molecules represented as binary fingerprints is calculated using the Tanimoto coefficient [44] which gives the measure of overlapping bits between the two molecules.

The Tanimoto similarity (S_{AB}) between two molecules A and B is given by:

$$S_{AB} = \frac{c}{a+b-c}$$

Where a and b represents the number of bits set to 1 for molecules A and B respectively and c represents the number of common bits set to 1 between them. The range of Tanimoto

coefficient is between 0 to 1 where 1 indicates the maximum similarity where as 0 indicates there is no similarity.

For example, the Tanimoto similarity between two molecules A and B represented as binary vectors of length 10 can be calculated as shown below :

$A =$ | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |

$B =$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

$$a = 7,\ b = 5\ \ and\ \ c = 5$$

$$S_{AB} = \frac{5}{7+5-5} = 0.71$$

Molecular similarity strategy is based on the evaluation of k-nearest neighbors (*k*NN) approach. Each compound in the evaluation set was compared to all the compounds in the training set and average similarity of the three most similar compounds were considered to predict toxicity class. In order to determine the best parameter for the similarity search, a leave-out-cross-validation was carried out on the training set with two different fingerprints separately.

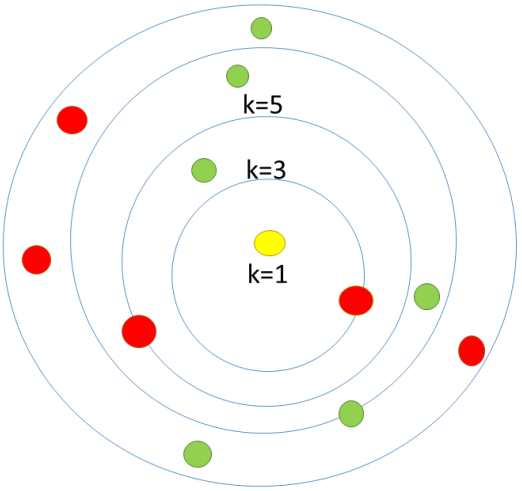


FIGURE 3.2: **Illustration of a *k*NN classification model** For k=1, the classification of the yellow query instance as a member of the red class; for k=3 it will be again assigned to the red class based on a 2-1 vote. However in cass of k=5, the nearest neighbours are both green, the model will classify it as a part of the green class with a 3-2 vote.

3.4.2 Fragmentation method

In this section, a new method to predict the toxic effects of the chemical compounds based on their local features (chemical fragments) has been proposed. The uniqueness of this method is the ability to extract correlated sets of chemical substructures and their possible association with certain toxicity class. This method is based on two steps: first fragmentation of the molecules in the data set using two different types of fragmentation methods and second, using propensity based statistical analysis of the fragments which are over represented in the toxic classes.

The complete data set molecules were fragmented using fragmentation method called ROTBONDS and RECAP[45]. The ROTBONDS fragmentation method cuts the molecule recursively at its rotatable bonds. So more number of rotatable bonds in a molecule results in more number fragments. Whereas, the RECAP is based on the rules of combinatorial chemistry and hence fragments created through this approach are easy to synthesis or joined (e.g through amide linkage, ester bonds, quaternary nitrogen bonds). The unique fragments created by ROTBOND and RECAP method were 37,251 and 45,712 respectively. After combining both ROTBONDS and RECAP fragments datasets and removing the duplicates, a total of 75540 fragments were created. Each fragment was mapped with an unique identifier and the method by it was created as well their presence in the associated toxic class molecules. For each of the fragments their occurance in individual toxic class was recorded using substructure search and propensity score as well as a confidence score were calculated using the formula described below

Let p1, p2, p3, p4, p5 and p6 represents respective toxic classes I, II, III, IV, V and VI

$p1_i$= occurrence of fragment (i) in toxic class I

similarly for $p2_i$, $p3_i$, $p4_i$, $p5_i$ and $p6_i$

Let d1, d2, d3, d4, d5 and d6 are distribution of fragment (i) in respective toxic classes I, II, III, IV, V and VI

Therefore equation (1),

$$d1_i = \frac{p1_i}{p2_i + p3_i + p4_i + p5_i + p6_i}$$

Similarly for $d2_i$, $d3_i$, $d4_i$, $d5_i$ and $d6_i$ are calculated for each fragment

Let T1, T2, T3, T4, T5 and T6 represents total number of fragments in toxic classes I, II, III, IV, V and VI respectively.

And dT1, dT2, dT3, dT4, dT5 and dT6 represents distribution of total number of fragments in toxic classes I, II, III, IV, V and VI respectively,

Therefore equation (2),

$$dT1 = \frac{T1}{T2 + T3 + T4 + T5 + T6}$$

Similarly dT2, dT3, dT4, dT5 and dT6 are calculated,

So the final confidence score is given by,

$$Pscore\ of\ Fragment(i)\ in\ toxic\ class\ I = \frac{equation(1)}{equation(2)}$$

Similarly for PScore for each fragments for individual classes were calculated using the above formula.

The Pscores were then normalized using Z score normalization. The Pscore of zero represents absence of the fragment in a particular class and 1 represents the maximum occurrence. The standard deviation curve for the Pscore was computed and threshold was considered to further filter the most relevant fragments in the data set. The total number of fragment was reduced to 32790. Finally, 7533 most toxic fragments representing the most toxic classes I, II and III with high PSscore and absent in the less toxic classes IV, V and VI were chosen and used as descriptor for the toxicity prediction model.

3.5 Results

Based on the comparison of the results of cross validation obtained using ECPF4, FP24 and FP24 combined with fragments separately for the prediction method, the best performance across all the toxicity classes was achieved by using ECFP4, followed by FP24 combined with fragments and FP24. However, for the most toxic classes (class I, II and III) were predicted more accurately by a consensus method taking FP24 and fragments into account, as shown in table 3.3.

3.6 Performance evaluation

The prediction method was evaluated on an external independent validation set. The validation set represents approximately 5 % of the dataset used for prediction method with similar

Sensitivity in %	**ECFP4**	**FP24**	**FP24 and fragments**
Over all	70	68.8	69.1
Toxicity class I	33	37.4	41.0
Toxicity class II	48	42	50.9
Toxicity class III	60	59	62.2
Toxicity class IV	81	80.6	80
Toxicity class V	58	56.5	56.1
Toxicity class VI	41	39.4	39.3

TABLE 3.3: **Performance of the prediction method in leave-one-out cross validation using different features** .

distribution of compounds for different toxicity classes. The method was evaluated on the following measures based on number of true positive (TP), true negative (TN), false negative (FN) and false positive (FP) as given below:

The true positive rate

$$Sensitivity = \frac{TP}{TP+FN}$$

The true negative rate

$$Specificity = \frac{TN}{TN+FP}.$$

The positive predictive value

$$Precision = \frac{TP}{TP+FP}.$$

Coverage rate is defined as the percentage of compounds for which a prediction could be made.

It was observed that consideration of the fragments increased the prediction rates for most toxic classes as observed in the cross-validation see table 3.3 , in particular for toxicity class I by 6 %, thus supporting the applicability of ensemble method. Furthermore, the coverage also improved by approximately 2 % as shown in the table below.

Performance measure in %	**FP24**	**FP24 and fragments**
Over all sensitivity	75.56	73.08
Sensitivity toxicity class 1	66.67	72.73
Sensitivity toxicity class 2	65.52	61.80
Sensitivity toxicity class 3	66.79	67.88
Specificity	95.11	94.62
Precision	75	73.50
Coverage	90.14	91.78

TABLE 3.4: **Performance of the external validation set** .

3.7 Comparison with other methods

To further evaluate the performance of the prediction method, it was further compared to commercial software TOPKAT (Accelrys Inc., USA) for oral rat LD $_{50}$ model and freely available software T.E.S.T (USA Environmental agency) oral rat LD $_{50}$ model based on nearest neighbor approach; over the external validation set as shown in table 3.5. Since, the consensus method including FP24 and fragments achieved the best performance in cross validation, this feature was considered for further comparison. The method presented in this chapter has outperformed both compared models. The overall sensitivity of developed prediction method is almost double the performance achieved by the compared models on the external validation set.

Performance measure in %	**FP24 and fragments**	**TOPCAT**	**T.E.S.T**
Over all sensitivity	73.08	44.8	46.27
Sensitivity toxicity class 1	72.73	0.00	0.00
Sensitivity toxicity class 2	61.80	1.15	22.37
Sensitivity toxicity class 3	67.88	29.43	23.33
Specificity	94.62	88.96	89:25
Precision	73.50	41.98	45.61
Coverage	91.78	89.40	78.64

TABLE 3.5: **Comparison of the prediction method with TOPKAT and T.E.S.T on the external validation set** .

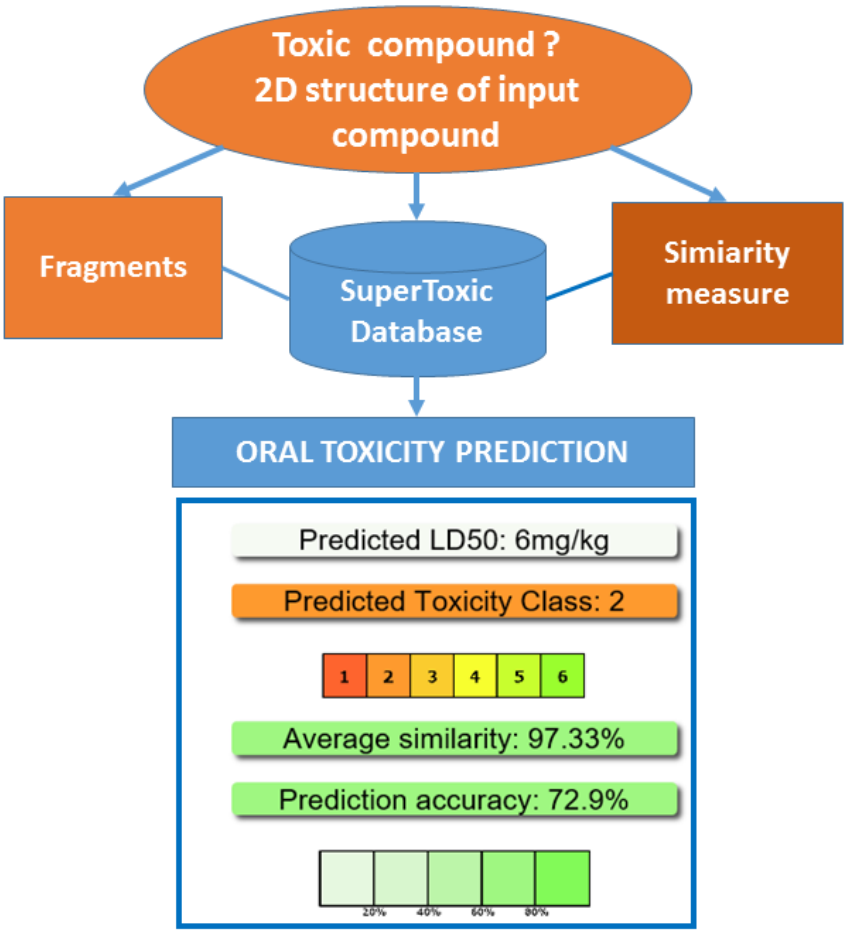


FIGURE 3.3: **ProTox web-server** Given the 2D coordinates of a compound, the web-server calculates the toxicity prediction of the compound using the ensemble approach and present the results with the predicted class, average similarity with the toxic compound in the training set as well as percentage of prediction accuracy. Additionally, if any toxic fragments are identified in the input compound, the fragment is reported with a confidence score.

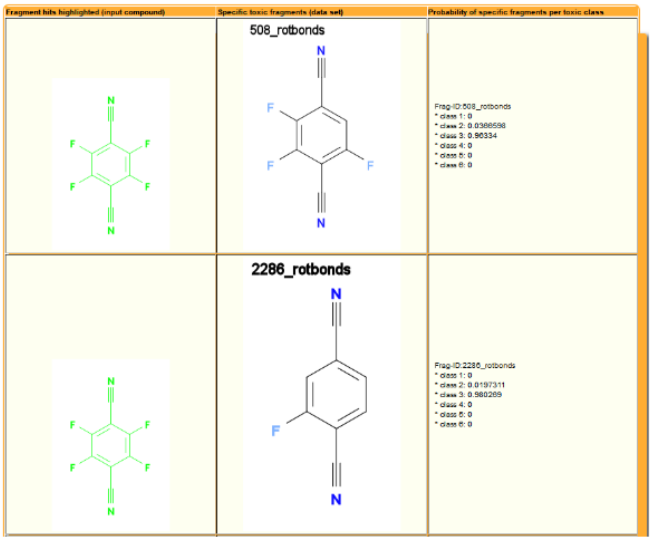


FIGURE 3.4: **ProTox web-server** Given the 2D coordinates of a compound, the web-server calculates the toxicity prediction of the compound and highlights the toxic fragments with individual confidence score.

3.8 Discussion

The study reported in this chapter establishes a novel method for prediction of rodent oral toxicity only available information on chemical data. Similarity method needs little information to formulate a reasonable query; specifically it is not necessary to know about the active confirmation of the query molecule as its based on the 2D coordinates of the molecules. Similarity based model can be really useful when little or no information about the target is known. Implementation of similarity method is computationally inexpensive and large data sets can be predicted and analyzed in less time. A major limitation of this approach is that prediction for each query (unknown) molecules depends largely depends on the activity and diversity of the structurally similar molecules present in its training set which account to the neighborhood behavior of the molecules. Though the molecules are compared using structural molecular fingerprints, however the scoring is based on the global similarity between the compared molecules. Often it is observed that the property of a chemical compounds is deeply connected with a local feature present in it. Hence, understanding of these local patterns is extremely important in order to understand the activity of a molecule. In this study, fragmentation approach was implemented inspired by the 'local feature based association' hypothesis. It was observed in this study, that addition of fragments descriptor improved the prediction for the most toxic classes such as class I by 5 % , class II by 9 % and class III by 3 % on cross-validation set as shown in table 3.3. Based on only individual fingerprints the prediction method achieved highest performance with ECFP4, followed by a combination of FP24 and fragments and only FP24 fingerprints individually has the least performance

on cross-validation set. However, the best performance on external validation was achieved by combination of FP24 and fragments. The ensemble method developed in this study was compared with the commercial software TOPKAT and publicly available QSAR based method T.E.S.T and has outperformed them in all evaluation measures on the external set as shown in table 3.3 .

3.9 Conclusion

In this study, an ensemble model based on similarity and fragments approach was developed to predict rodent oral toxicity of the chemicals by using concatenated 'FP24' fingerprints and statistically analyzed toxic fragments. The similarity approach is based on the neighborhood behavior of chemical compound and representation of their structural features using molecular fingerprints. In the fragment based approach, each molecular structure was compared to statistically analyzed fragments by using substructure search that represents the presence or absence of particular toxic substructures in the molecules with a confidence score. The application of this ensemble approach has the potential to achieve a classification model with high prediction accuracy as well as prediction confidence. The major advantage of this approach is the capability to incorporate addition of new toxicity data easily and the model can be extended to other toxic end points.

3.10 Application of the *in silico* prediction method in the development of natural product database.

Natural products (NPs) plays an important role in the drug discovery pipeline. The constant emergence of new natural product chemotypes with interesting structures and biological activities leads to potential sub-librtary generation for target based screening. It has been observed that many topological pharmacophore patterns are common between natural products and commercial drugs. Most recently, the novel prize 2015 in medicine was given for discovery of natural products artemisnin and avermectin [46]. This supports the fact that a better understanding of the specific physicochemical and structural versatility of natural products is important in the modern drug discovery development. Due to growing interest in natural products and its applications in the field of drug discovery, this was indeed important to create a database of natural product which is publicly available. NPs are often associated with their complicated ring structures containing more than two rings often termed as macrocycles (ring with more than 12 atoms). This is a special feature of NPs which help them reorganize themselves structurally such that the important functional groups can strongly interact with the target proteins. This is mainly done to reduce the entropic loss associated

with the binding of ligands (NPs) with the proteins[47]. Macrocyclic NPs like erythromycin, rampamycin, tacrolimus have physicochemical properties like lipophilicity, metabolic stability, increased solubility and bioavailability[48] which make them desirable drug candidates. Drugs like topotecan and irinitecan which has been approved for cancer therapy, has been inspired from camptothecin isolated from the bark and stem of *Camptotheca acuminata*[49]. Camptothecin was initially found to inhibit the DNA enzyme topoisomerase I. However, due to low solubility and high adverse drug reactions during clinical trails[50], it has to be modified to its analogs topotecan and irinitecan.

At the same time, NPs are also known to be toxic and produce adverse effects on cellular or system level. Alpha-amanitin, a well known toxic peptide produced by *Amanita* mushrooms can lead to fatal liver and kidney disorders[51]. Though the potentiality of NPs as drugs has no doubt and well established in the area of drug design, however NPs are not excluded from the curse of adverse drugs effects. Therefore, it is also important to study the toxicity profile of NPs which is been considered as an therapeutic agent.

Even though there is a growing interest of NPs, there are only few NPs based database available for the scientific community. Most of the well-known NPs based database such as the Dictionary of Natural Products[52] and Natural Product Alert (`https://www.napralert.org/`) are either commercial or freely available only with restricted information. There was certainly a need of a NPs based database that is available freely to the scientific community, containing information on various level like classification, toxicity class, mechanism of action and pathways information. The knowledge based pool of Super Natural II database was created keeping all the above mentioned factors. Super Natural II is the first publicly available database of natural compounds consisting of 326,00 molecules.

The total number of compounds present in the Super Natural II database is 326,000. These compounds were obtained from multiple data sources including; databases, literature review, and expert knowledge as cited in the Super Natural II website . The Kyoto Encyclopaedia of Genes and Genomes (KEGG)[53] is a popular metabolic database. This database is used to generate the metabolic pathways for the natural compounds in the Super Natural II database. Cross-references between resources are provided in the database either internally or externally via identifier mapping.

Some of the database sources which have been used in the creation of Super Natural II database is listed below

Source	Number of compounds
Ambinter	48649
Analyticon Discovery	32641
Cyano Biotech GmbH	32
HMDB	4744
Indofine	572
InterBioScreen	84265
Iris	959
KEGG	2324
MedChemLabs	128
MetaCyc	350
Microsource	501
Molecular Diversity Preservation International	26639
Nubbe Natural Products	581
Princeton Biomolecular	27394
Selleckchem	178
Sigma	364
Specs	3498
TCM Database	23943
TimTec	70
UEFS Natural Products	117
Universal Natural Products Database	166097

3.10.1 Toxicity prediction

The toxicity related to NPs which are used as medicine is usually not well established. There is a lack of standardized reporting system as well as limited interest in the toxicity profiles of already known drugs from natural sources. In the year 2013, a study reported to the American Association of Poison Control Centers, almost 38,955 natural product-related exposures were reported[54,55]. Most of the NPs are not always marketed in their pure form , but in combination with more than one NPs in mixture. This certainly adds to the complexity to identify a cause and -effect relationship between an individual NPs and specific toxicity. Adverse effects of NPs have been associated with nephrolithiasis, rhabdomyolysis, hepatorenal syndrome and many more[55]. For instance, daily intake of creatine monohydrate has been associated with acute folcal interstitial nephritis with tubular injury[56]. On the other hand, some toxic profiles of NPs are also important for killing tumorous or infectious cells, in case of NPs with anticancer properties. It should be noted that some NPs are toxic in nature and hence, there is a need to discriminate the non-toxic NPs and toxic NPs. The toxicity profiles of any

compound is dependent on the dose and duration as well as physiological state of the system administered; and have to be studied in details and cannot be generalized. Nevertheless, having some first hand information of the toxicity profile based on its structural feature will lead to be a preventive measure. The toxicity prediction is categorized into six major classes, ranging from class I to class VI based on the lethal dose (LD_{50}) values in mg/kg body weight in rodents. The toxicity prediction method involved in this section is based on ProTox[57].The total number of compounds for which toxicity information is available is 170000 compounds, the unpredicted compounds were mostly due to out of prediction range of the method and so reported as zero. All the NPs in the database has a link connected to Protox web-server[57], which can address the toxicity prediction in details. As example of compounds with toxicity class information in Super Naturall II database is shown in the figure 3.10.1

Toxicity class	**number of compounds**
I	2392
II	10579
III	11969
IV	60821
V	59892
VI	20716

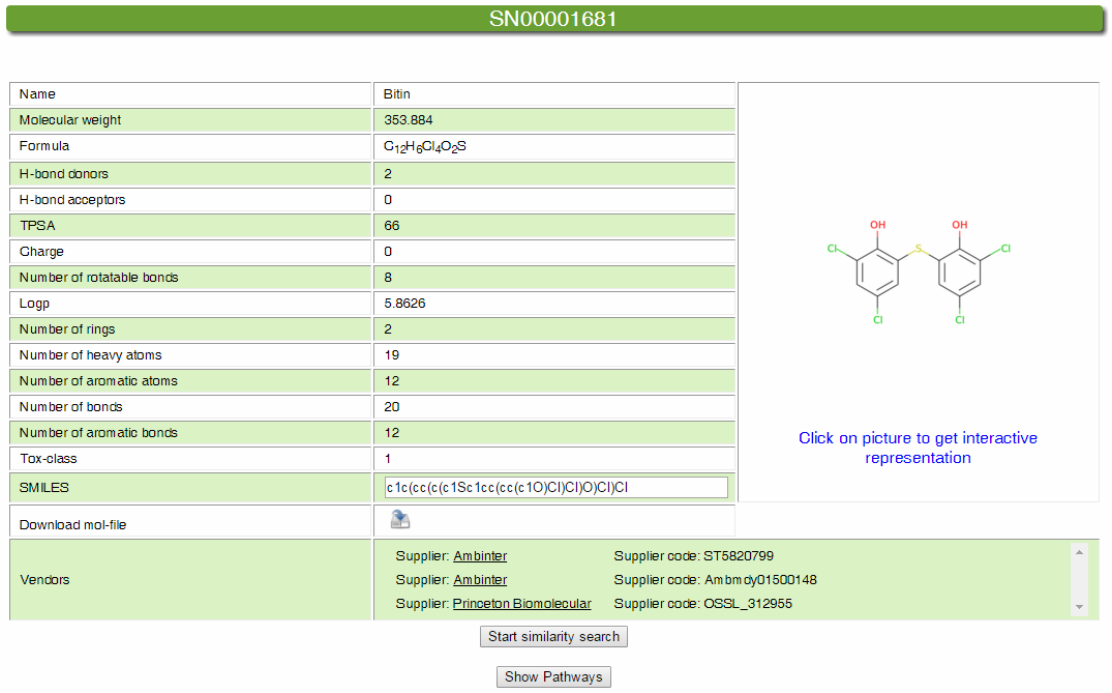


FIGURE 3.5: **Super Natural database** Given the 2D coordinates of a compound, the calculates the toxicity prediction of the compound using the ensemble approach and present the results with the predicted class, as well as properties of the molecule and further possibility to search pathways.

3.10.2 Graphical user interface and information retrieval

In this section, the different search options available via Super Natural II database are described. To help the user, a web-based search tool including a molecular structure sketcher interface was designed. Users can also search the database using properties, toxicity class, or by combination of criteria. Furthermore, the integrated PubChem[58] search for compound name is implemented. For every search results the main properties and the chemical structure of the compound, as well as a link to download the structure file in the Mol2 format is implemented.

Apart from toxicity prediction, some of the important search options are listed below:

1. **Similarity search**: To compare two compounds on the basis of their structural similarity, a molecular fingerprint based similarity search was implemented using Open Babel library. The two-dimensional similarity of the compound were given a Tanimoto score. It compares the structural similarity between the query molecule and the database entries using a concatenated fingerprint of the 'FP2' and 'FP4' Open Babel

fingerprints. Pre-computed fingerprints for all database entries are stored as blob objects in the MySQL-database. For any input query structure it is calculated during the search. The top 15 similar compounds are retrieved as result.

2. **Mechanism of action**: The elucidation of precise mechanism of action of NPs remains a considerable challenge in drug discovery. Target identification and mechanism of action can be obtained by direct biochemical essays, genetic interactions or computational inference. Information of mechanism of action of unknown compounds greatly facilitates the use of such compounds as a starting point for investigation of fundamental biological processes. Some NPs and other compounds have already known mechanism of action and target information. This information was used to predict the mechanism of action of unknown NPs which share structural similarity with the known compounds. The drug-target information for more than 195 000 pharmacologically known compounds were extracted from SuperTarget[59]. These drugs were compared with all the compounds present at the Super Natural database using structural similarity score (Tanimoto coefficient) of 0.8 and above. These information were stored in the database using an unique identifier. SuperPred web-server was used to predict the targets for the natural product compounds, which has a prediction accuracy of 75.1 percent[60].

3. **Pathways**: In this section, the information from KEGG pathway were mapped for the natural compounds present in the Super Natural II database. The NPs were mapped to potential targets with respect to their similarity to the reference drugs (as explained under 'mechanism of action' section), to display pathway maps. Taking an example for a search considering 'Homo sapiens' as species information and 'Apoptosis' as pathway, the information can be obtained as a pathway map with all known targets highlighted. A curser-over on the target will display drugs action on the target. The number of drugs is limited to 15 hits with a similar NPs and their related similarity score. The database currently considers metabolic pathways of Homo sapiens, bacteria and fungi.

3.11 Availability

The method is made publicly assessable via a web-server called 'ProTox' (Prediction of Rodent Oral Toxicity (`http:tox.charite.de`). It is the first freely available method based on chemical similarity and the identification of toxic fragments for *insilico* prediction of rodent oral toxicity. This method demonstrates good performance in comparison to available QSAR-based methods.

The Super Natural II database is a freely available knowledge resource of natural compounds with different embedded search options. The compounds present in database comprise of

diverse chemotypes with a wide range of biological as well as pharmacological activities. Additionally, toxicity profiles of the natural products are provided. Therefore, it is believed that the database will be effective and beneficial in the studies involving metabolomics, virtual screening and design of novel compounds. Since its publication, Super Natural database II database has been used in many researches for novel discoveries[61,62] as well several analysis purposes. Super Natural II database is available at (http://bioinf-applied.charite.de/supernatural_new/index.php)

Chapter 4

Development of binary classifiers for predictions of compounds active in toxicological pathways

Several approved drugs are withdrawn from the market for safety reasons and there are many reasons including metabolism and different mechanism that can cause toxicity[63] . The current improvement in the field of predictive science related to toxicity prediction led to the development of several computational methods. These methods uses the molecular descriptors, biological data and chemical structures and analyze them using a statistical analysis; represents them in a mathematical equation. These mathematical models describes the relationship between the structures and activity and predicts the toxic effects[65]. This kind of approach is an unbiased way to asses the data to generate relationships and predict their toxic outcomes. Such approaches has the capability to discover potentially new SARs (Structure Activity Relationships) and can lead to new ideas in assessment of mechanisms by which chemical interact with biological systems. However, the results are highly dependent on the quality of the chemical space used to build the model, so a careful validation is important for the effective used of these approaches.

In this chapter, different *in silico* approaches used to predict the outcomes of the compounds that have the potential to interact with the molecular targets active in toxicological pathways are discussed. The data used in this study was obtained from the Tox21 Challenge platform (http://www.epa.gov/chemical-research/toxicology-testing-21st-century-tox21) provided by the National Institute of Health, USA (NIH). Tox21 challenge data was created using standard assays and experimental conditions and therefore serves as the 'gold standard' data to asses and compare computational prediction methods.

In the first section of this chapter, a similarity-based fingerprint approach is discussed, which is inspired from the method explained in the previous chapter. This method was used to predict the activity of the compounds in the Tox21 challenge and has been ranked 8th out of 30 models, 9th out of 32 models and 9th out of 35 models for the targets SR-HSE (Heat shock factor response element), ER-LBD (Estrogen receptor alpha) and NR-AhR (Aryl hydrocarbon receptor) respectively, submitted to the challenge (https://tripod.nih.gov/tox21/challenge/leaderboard.jsp).

In the second section, different machine learning models are discussed which were developed using the same Tox21 challenge data and their performance as compared to the top models in the Tox21 challenge for the three targets. As well as their better performance compared to the method similarity-based fingerprint approach described in the first section. Additionally, taking example of one of the targets, an analysis of the active chemical space was done to understand which machine learning models can correctly predict which chemotypes with highest confidence scores.

4.1 Introduction

In the year 2008, the U.S National Institutes of Health (NIH) and the U.S environmental Protection Agency (EPA), later joined by the U.S Food and Drug Administration (FDA) came together to establish a platform for the future of toxicity testing. With the vision to transform toxicology into predictive science, they came up with the initiative 'The U.S Toxicology in the 21 st Century (Tox21)'[64]. The main objective of this initiative is to shift *'the traditional experimental animal toxicology studies to one based on target-specific, mechanism-based, biological observations largely obtained using in vivo assays'*[43] . Through this consortium, a platform 'the Tox21 Data Challenge' was organized for computational toxicologists to develop and validate their predictive models. The consortium screened a large library of compounds, including chemicals and drugs involved in different pathways and responsible for eliciting toxic effects.

4.2 Targets / Assays

In this study targets from two different pathways namely: nuclear receptor signaling pathways (AhR and ER-LBD) and stress response pathways (HSE) are considered. The toxicity profiles of these targets are well established and are taken from the U.S Tox21 program.

The Aryl hydrocarbon receptor (AhR) is a ligand activated transcription factor, involved in the regulation of biological responses to planar aromatic (aryl) hydrocarbons. Side effects

associated with this target results in a adaptive responses to environmental changes by induction of metabolising enzymes (CYP 1A 1 and CYP 1B1), producing toxic metabolites. It is also known to interact with other nuclear receptors[65] .

Estrogen Receptor (ER) is a nuclear hormone receptor which plays an important role in metabolic homoeostasis and reproduction. ER is responsible for the regulation of genes involved in growth and development of various tissues. Many chemicals like diethystilbestrol, bisphenol A are known to binds to ER- receptor and inhibits growth. Tamoxifen is a drug which is involved in fatty liver disease as well as liver toxicity is known to bind to ER receptor and inhibit the gene transcription process[65] .

Many chemicals under environmental and physiological stress state may activate heat shock /unfolded protein response. Therefore Tox21 program developed a HSE-bla Hella cells based assays to identify compounds that are active in HSR signalling pathway[65] .

4.3 Data preparation and processing

The publicly available Tox21 data challenge 2014 10k library was taken from the Tox21 challenge website containing the 2D structure of the molecules. The standardization of the chemical structures was done using the Instant Jchem software (version 6.2 , ChemAxon). The standardization protocol includes the following step:

1. Removal of water molecules.

2. Aromatization of all the structures.

3. Transformation of adjacent positive and negative charges into double or negative bonds, wherever applicable.

4. Removal of explicit hydrogen.

InChIKeys for the molecules were calculated using RDKit nodes in KNIME; this was done to keep an account of the duplicate structures and filter them. In case identical molecules for a given target had two different activities value (i,e one being active and other inactive), they were marked ambiguous and removed from the training set. After the standardization protocol and duplicate removal, the final training set and external validation set had similar class distribution as shown in tables 4.1 and 4.2 .

Target	Total	actives	inactives	Ratio (actives/inactives)
AhR	6901	769	6132	0.125
ER-LBD	6801	346	6455	0.053
HSE	7328	308	7019	0.043

TABLE 4.1: **Training set class distribution**. The table shows the total number of compounds as well as number of actives and inactives compounds for each target.

Target	Total	actives	inactives	Ratio (actives/inactives)
AhR	610	73	537	0.135
ER-LBD	600	20	580	0.034
HSE	610	23	588	0.039

TABLE 4.2: **External test set class distribution**. The table shows the total number of compounds as well as number of actives and inactives compounds for each target.

4.4 Molecular fingerprints

Four different fingerprints were used in this prediction method.

4.4.1 Substructure fingerprints

Substructure fingerprints are based on *structural keys* which were primarily designed for high speed screening of chemical databases. The structural keys are commonly represented as a boolean array or bitmaps, in which each bit represents presence or absence of a specific structural feature. These patterns include the elements, important electronic configuration (such as 'sp^3 or triple-bond nitrogen), common functional groups (such as alcohols, amines, hydrocarbons), ring systems (such as cyclohexane, pyridine or naphthalene), functional group of special importance (such as organo-metallic group or drugs like steroids based structures). The public Molecular ACCess System (MACCS) structural keys are 166 predefined substructures defined as SMARTS based on dictionary-based class of fingerprints[66] . These fingerprints were created using the RDKit and CDK nodes in KNIME.

4.4.2 Circular fingerprints

This fingerprint was explained in the section 3.3.2

4.4.3 Estate fingerprints

The Estate fingerprint is based on the the concept of Electrotopological State Indices for atom types generated from a combination of electronic, topological and valence state information of an atom as described by Hall and Kier[67]. The fingerprint contains 79 bits using the E-state based concept fragments. These fingerprints were created using the cdk node in KNIME .

4.4.4 Toxicity fingerprints

Toxicity based fingerprint 'ToxPrint' used in this study includes a public set of chemotypes encoding the generic geno-carcinogenic substructures (https://toxprint.org/).

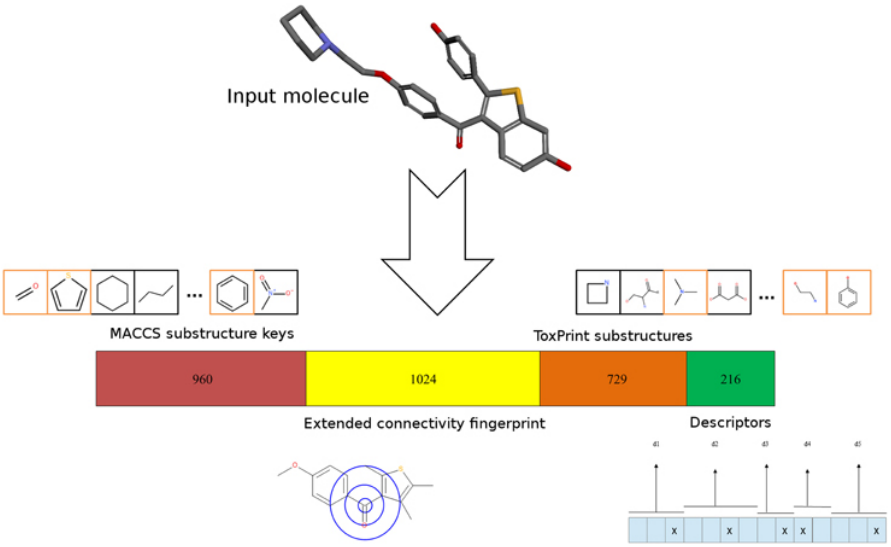


FIGURE 4.1: **Concatenated fingerprint** A combination of sub structural fingerprint (MACCS), toxic alerts based fingerprint (ToxPrint), circular fingerprint (ECFP4) and property-based binary fingerprint[64]

Fingerprints	**Designed descriptors**	Encoding	**Length**
MAACS	structural fragments	one to one matching of bits position and fragments keys	166 or 960 bits
ECFP	extended graph connectivity	mapping of hash function to virtual feature space	infinite
FP2 or FP4	paths or subgraph	mapping of hash function to fixed length	user-defined (1024 or 2024 bits)
Estate	valence state information of atom	mapping of hash function to fixed length	fixed length
ToxPrint	substructures	one to one matching of bits position and substructure keys	fixed length

TABLE 4.3: **Types of fingerprints and their encoding description**. Five different fingerprints used in this study and their respective encoding parameters.

4.5 Molecular descriptors

Molecular descriptors encoding the physico-chemical properties of the molecules for the data set were calculated using the the RDKit descriptor calculation node in KNIME[68]. A total of 43 descriptors were calculated and their individual scores were normalized using Z-score normalization. The descriptors with missing values were removed. Additionally, by plotting the 'principal components', descriptors with low variance to both active and inactive class were removed. Finally, 13 descriptors were selected and used in this study as presented in the table 4.4 . The SlogP descriptors account for the hydrophobic and hydrophilic effect[69]. Chi indexes takes into consideration of valence value to encode sigma, pi and lone pair interaction[70]. Kappa shape descriptors compares the molecule with extreme shape for the number of atoms[70], the applications of Kappa shape descriptors in the field of *in silico* predictions has been extrusive[69] .

Descriptor name	Description
SlogP	Log of the octanol/water partition coefficient
Chi0v	Atomic valence connectivity index (order 0)
Chi1v	Atomic valence connectivity index (order 1)
Chi2v	Atomic valence connectivity index (order 2)
Chi3v	Atomic valence connectivity index (order 3)
Chi4v	Atomic valence connectivity index (order 4)
Chi1n	Simple molecular connectivity index for path (order 1)
Chi2n	Simple molecular connectivity index for path (order 2)
Chi3n	Simple molecular connectivity index for path (order 3)
Chi4n	Simple molecular connectivity index for path (order 4)
kappa1	Kappa index for 1 bonded fragment
kappa2	Kappa index for 2 bonded fragment
kappa3	Kappa index for 3 bonded fragment

TABLE 4.4: **Molecular descriptors used in the method in combination with fingerprints** .

4.6 Methods

In this section, different methods developed for prediction of the binary classes of Tox21 data are described. Firstly, a similarity based approach in combination with Naive Bayes fingerprints learner is reported, followed by different machine learning algorithms such as Naive Bayes (NB), Random Forest (RF) and Probabilistic Neural Network (PNN).

All the methods takes 2d structure of chemical compounds as input and predicts the outcomes with confidence score for the compound to be active or inactive for the given target.

4.6.1 Similarity-based fingerprint method

Similarity based method is based on the similar strategy as explained in the section 3.4.1. Additionally, three Tanimoto coefficients (Tc) were computed, the maximum Tc to actives in

the training set (T_1), the average Tc to actives in the training set (T_2), and the maximum Tc to all inactives in the training set (T_3).

The Naive Bayes fingerprint learner is a variant of a Naive Bayes for fingerprints columns. This is not a model based on Naive Bayes algorithm, however this variant implements the Naive Bayes like algorithm that incorporates sparsely occupied bits and unbalance class distributions

The consensus score from both the method was considered after normalization of all the scores using Z-score normalization followed by a Gaussian distribution and normalization implemented in KNIME. The scores were obtained in a range of 0 to 1. The combination of each scores were considered and ranked and the highest scores were considered as the best score[64] .

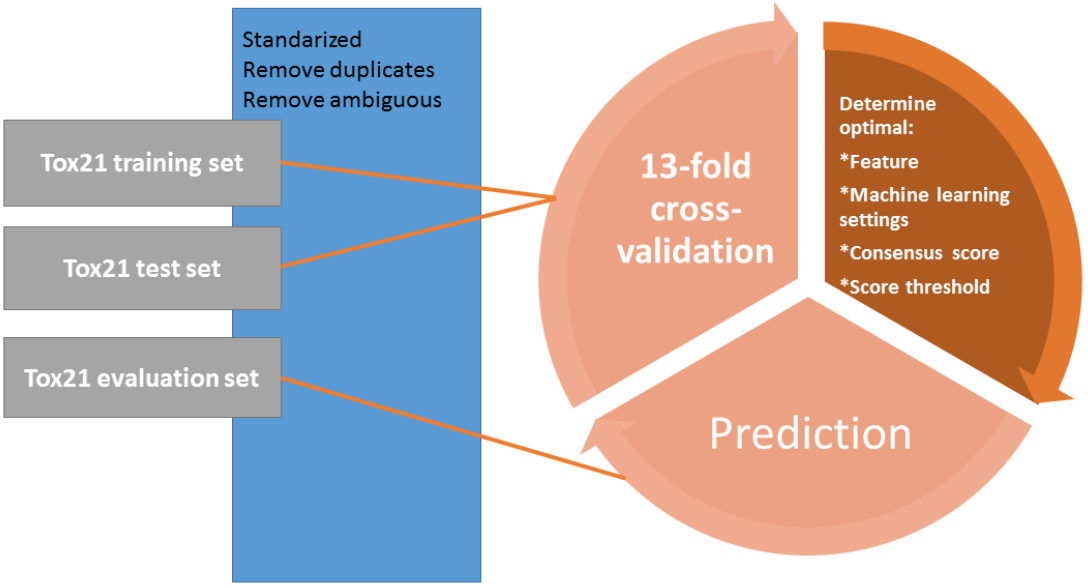


FIGURE 4.2: **Workflow of the similarity-based fingerprint method** Schematic representation of the similarity based method models used to predict the outcomes of Tox21 data

4.6.2 Naive Bayes

The Naive Bayes classifier is based on the assumption of the Bayesian theorem of conditional probability[71] . In order to classify data point, NB assumes an independence of features and calculates the overall probability for each class. In the study, NB estimates the probability for a given chemical structure to be active based on the presence or absence of structural features for a given target. Based on these individuals features, NB classifies a compound as active if the probability for being active exceeds the probability for being non-active. NB classifier was implemented using the Naive Bayes Learner and Predictor nodes in KNIME[71] .

The Learner node takes the training data as well as the features and parameters for the model. The maximum number of unique nominal values per attribute was set to 20. The predictor node takes the model and test data and as output classifies the test data with an individual class and score[71].

The model was trained using different molecular fingerprints and molecular descriptors to obtained the optimal prediction parameters.

4.6.3 Random Forest

Random Forest classification which is based on decision trees was developed by Breiman[72]. In this classifier each tree is independently constructed and each node is split using the best among the subset of predictors randomly chosen as node. That is, RF is capable of describing the relationship between independent features such as fingerprints, molecular descriptors with dependent variables such as activity and toxicity. The RF grows collection of trees often termed as forest, and uses these trees for classifying a data point into one of the classes. The type of randomness used in our model to make the the classification trees grown in the forest are dissimilar and uncorrelated from each other and is based on random selection of input variables. The split criterion Gini which has been accepted as best choice previously[73] was chosen and obtained best performance for the target AhR. However, for ER-LBD and HSE, information gain ratio as split criterion gave best results. The number of trees in the forest was limited to 1000 and a data sample of ratio of (80:20) for AhR and (70:30) for ER-LBD and HSE was chosen with replacement for each tree, this step is similar to bootstrapping. This was done to achieve low error rate of convergence. Furthermore, a square root function was used for attribute sampling and different sets of attributes were chosen for all the trees. The predictor node predicts the class for each compound in the test data based by the majority vote classification trees in the forest, each compound has an overall prediction score and individual class confidences. This model was implemented using the Tree Ensemble Learner and Predictor nodes in KNIME.

4.6.4 Probabilistic Neural Network

The Probabilistic Neural Network is based on a statistical algorithm 'Kernel discriminant analysis'[74] . PNN is a four layer, feed-forward neural network that is widely used in the area of pattern recognition, nonlinear mapping and estimation of probability of class membership and likelihood ratios[75]. The first layer which is also the input layer consists of sets of measurements. The pattern layer which is the second layer consists of the Gaussian function uses the given set of data points as centers. The summation layers or the third layer performs

an average operation of the outputs from the second layer for each class. The fourth layer that is the output layer predicts the class based on votes from the largest values[76]. Similarly, like NB and RF, PNN was also implemented using KNIME. The PNN learner node takes the numerical data as input and PNN predictor node predicts the test data with a confidence score and class. In this model, all the parameters value were kept as default except for that the maximum number of Epochs was set to 42 to reduce the computational time complexity.

4.7 Constructions of the machine learning models

The Tox21 dataset for the respective targets was separated into 80 % training set and 20 % internal test set. A stratified sampling based random partitioning was done in order to keep the similar class distribution in both the sets. This was done to tune the model parameters and to avoid any random prediction.The data set used to train the model was highly imbalanced having active class (minority) and inactive class (majority). Due to lack of feasibility to obtain more data for the minority class in case of most targets, a stratified random sampling was applied to handle the classification accuracy error due to imbalanced data. Additionally, a set of 647 chemical structures was considered as external validation set. In order to optimize the best descriptors for individual models, four different fingerprints and rationally selected physico-chemical properties based molecular descriptors were chosen to represent the chemical structures. Using these descriptors individually and in combination, three different machine learning models were developed.

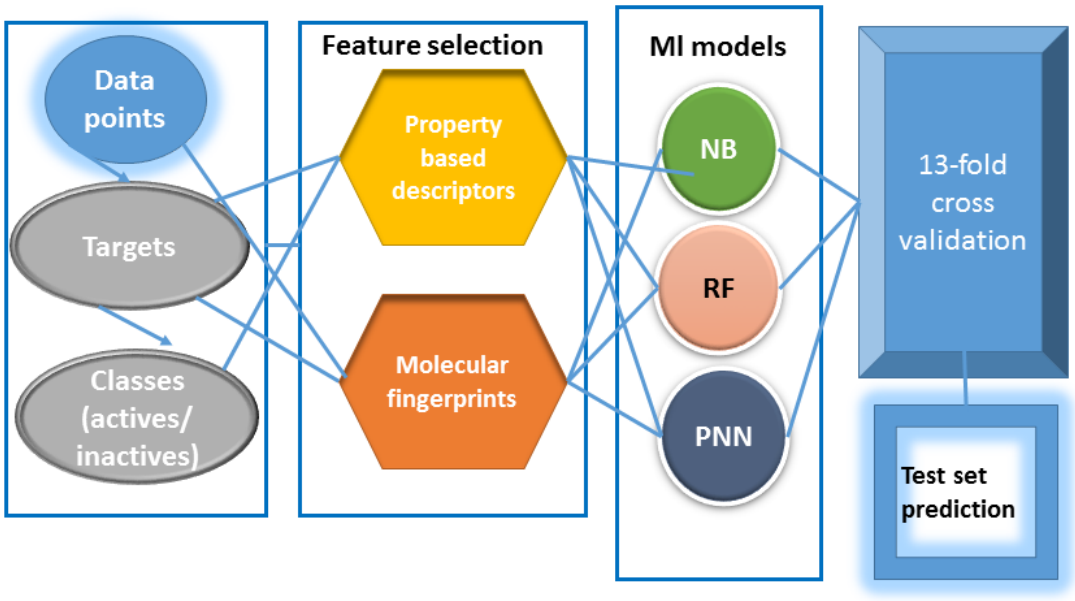


FIGURE 4.3: **Workflow of the machine learning methods** Schematic representation of three different algorithms developed for the predictions of active compounds in three different target classes respectively

4.8 Performance evaluation

In this study to evaluate the models performance, the area under the receiver operating characteristic (ROC) curve, also known as AUC was considered along with balance accuracy. ROC curve plots the true positive rate against the false positive rate and is traditionally used for binary classifiers[77]. A perfect model is believed to have AUC of 1 and a random model will have a value of 0.5.

$$ROC-curve = \frac{Sensitivity}{1-Specificity}$$

Additionally, in order to avoid inflated performance estimates on the imbalanced data balanced accuracy as another performance measure was considered. Balanced accuracy is defined as the arithmetic mean of sensitivity and specificity[78].

$$Balanced-accuracy = \frac{Sensitivity+Specificity}{2}$$

The models were validated based on two validation approaches, 13-fold cross-validation and independent external validation. For 13-fold cross-validation the data set was split into 13

roughly equal-sized parts, and then the model was fit into 12 parts of the data and the error rate was calculated in the other part. The process is repeated 13 times so that each of the part can be predicted as the validation set.

AUC and the cross-validation protocol was implemented in KNIME using the ROC curve node and cross-validation meta node respectively.

4.9 Results

In this section the results of the different prediction algorithms are explained in detail. In case of similarity based-fingerprint method, concatenated fingerprint was implemented and for the machine learning methods four different fingerprints ECFP4, MAACS, ToxPrint and Estate fingerprints in combination with different molecular descriptors were used. In this study it was observed that all the classifiers achieved almost prediction accuracies of 80 % and above. Due to the imbalanced data set, accuracy cannot reflect the performance of the models. The models were additionally evaluated on the ROC-AUCs which is more accurate performance quality criterion for the models. Based on the results obtained from 13-fold cross-validation and external validation, RF based models achieved the best performance for all the three targets followed by NB and PNN.

For all the targets it was observed that RF model with MACCS fingerprints as feature exhibit the best performance. Only in case of AhR, ToxPrint fingerprints also performed equally well with an AUC value of 0.89 and 0.88 for the external validation and cross-validation respectively (see tables 4.9 and 4.10) which is comparable to the performance of the similarity-based fingerprint method. The total number of actives as well as molecules for AhR target was relatively large as compared to ER-LBD and HSE; this confirms that the size of the training set as well as the ratio between active and inactive molecules is an important factor contributing to its superior performance for AhR (see tables 4.1 and 4.2).

In the following sections a comparison of different molecular fingerprints and thei combination with the molecular property based descriptors for different models on cross validation as well as external validation set are provided in details.

4.9.1 Results: Similarity-based fingerprint method

In this method all the models based on individual fingerprints (ECFP4, MACCS and ToxPrint) showed good performance with an AUC mostly above 0.75. However, the best performance was achieved by a concatenation of all three fingerprints with a property-based-fingerprint encoding the topology of the molecules. This method was used in the Tox21 challenge to

submit results. The best results were achieved in the Tox21 challenge using this method is for the target ER-LBD[64] .

Method	Descriptors	NR-AhR	ER-LBD	HSE
Similarity based fingerprint	concatenated fingerprints	0.90	0.86	0.80

TABLE 4.5: **Cross validation results for Similarity based prediction**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

Method	Descriptors	NR-AhR	ER-LBD	HSE
Similarity based fingerprint	concatenated fingerprints	0.89	0.79	0.85

TABLE 4.6: **External validation results for Similarity based prediction**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

4.9.2 Results: Naive Bayes

In case of NB classifier, MACCS fingerprints in combination with molecular property based descriptor and ToxPrint fingerprints performed comparatively better for AhR with AUC values of 0.82 and 0.83 (cross-validation) and 0.84 and 0.82 (external validation) respectively as shown in the tables 4.7 and 4.8. On the other hand, the performance for the targets ER-LBD and HSE were poor with an AUC value below 0.75 both in cross-validation and external validation.

Method	Descriptors	NR-AhR	ER-LBD	HSE
NB	MACCS	0.83	0.73	0.67
NB	MACCS + descriptors	0.82	0.73	0.63
NB	EState	0.81	0.73	0.71
NB	EState + descriptors	0.79	0.77	0.69
NB	ECFP4	0.77	0.76	0.70
NB	ECFP4 + descriptors	0.78	0.77	0.69
NB	ToxPrint	0.83	0.71	0.68
NB	ToxPrint + descriptors	0.80	0.72	0.70

TABLE 4.7: **Cross validation results for Naive Bayes model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

Method	Descriptors	NR-AhR	ER-LBD	HSE
NB	MACCS	0.82	0.69	0.79
NB	MACCS + descriptors	0.84	0.70	0.77
NB	EState	0.79	0.67	0.72
NB	EState + descriptors	0.78	0.67	0.60
NB	ECFP4	0.77	0.71	0.76
NB	ECFP4 + descriptors	0.78	0.71	0.74
NB	ToxPrint	0.82	0.63	0.63
NB	ToxPrint + descriptors	0.83	0.60	0.60

TABLE 4.8: **External validation results for Naive Bayes model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

4.9.3 Results: Random Forest

RF classifier for the target AhR showed a good performance with MACCS, ECFP4 and ToxPrint fingerprints with an AUC values above 0.88 on the 13-fold cross-validation as well as on the external validation. The highest AUC scores of 0.90 and 0.91 (cross validation) and 0.90 and 0.87 (external validation) were achieved with MACCS fingerprint individually and MACCS fingerprints in combination with molecular property based descriptors, respectively. The combination of descriptors did not improve the AUC value of the eternal set in this case. Similarly, MACCS fingerprints based RF model achieved highest performance for ER-LBD and HSE with AUC values of 0.83 and 0.80 (cross-validation) and 0.81 and 0.86 (external validation) respectively as shown in the tables 4.9 and 4.10 .

Method	Descriptors	NR-AhR	ER-LBD	HSE
RF	MACCS	0.90	0.83	0.78
RF	MACCS + descriptors	0.91	0.86	0.80
RF	EState	0.77	0.59	0.66
RF	EState + descriptors	0.79	0.55	0.67
RF	ECFP4	0.87	0.82	0.77
RF	ECFP4 + descriptors	0.89	0.80	0.78
RF	ToxPrint	0.88	0.80	0.78
RF	ToxPrint + descriptors	0.88	0.85	0.80

TABLE 4.9: **Cross validation results for Random Forest model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

Method	Descriptors	NR-AhR	ER-LBD	HSE
RF	MACCS	0.90	0.81	0.86
RF	MACCS + descriptors	0.87	0.77	0.81
RF	EState	0.78	0.51	0.87
RF	EState + descriptors	0.76	0.74	0.73
RF	ECFP4	0.87	0.78	0.81
RF	ECFP4 + descriptors	0.88	0.78	0.83
RF	ToxPrint	0.89	0.71	0.70
RF	ToxPrint + descriptors	0.74	0.62	0.69

TABLE 4.10: **External validation results for Random Forest model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

4.9.4 Results: Probabilistic Neural Network

PNN based classifier classifier performed comparatively better for AhR with an AUC score of 0.84 (cross-validation) and 0.85 (external validation) with ECFP4 as the best feature. However, the performance for the targets ER-LBD and HSE was relatively poor and all the AUC values were mostly less than 0.80 for both cross-validation and external validation as shown in tables 4.11 and 4.12 .

Method	Descriptors	NR-AhR	ER-LBD	HSE
PNN	MACCS	0.90	0.83	0.78
PNN	MACCS + descriptors	0.91	0.86	0.80
PNN	EState	0.77	0.59	0.66
PNN	EState + descriptors	0.79	0.55	0.67
PNN	ECFP4	0.87	0.82	0.77
PNN	ECFP4 + descriptors	0.89	0.80	0.78
PNN	ToxPrint	0.88	0.80	0.78
PNN	ToxPrint + descriptors	0.88	0.85	0.80

TABLE 4.11: **Cross validation results for Probablistic Neural Network model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

Method	Descriptors	NR-AhR	ER-LBD	HSE
PNN	MACCS	0.81	0.69	0.77
PNN	MACCS + descriptors	0.82	0.76	0.72
PNN	EState	0.78	0.68	0.76
PNN	EState + descriptors	0.84	0.77	0.53
PNN	ECFP4	0.85	0.69	0.70
PNN	ECFP4 + descriptors	0.82	0.75	0.67
PNN	ToxPrint	0.82	0.69	0.67
PNN	ToxPrint + descriptors	0.83	0.74	0.60

TABLE 4.12: **External validation results for Probablistic Neural Network**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

4.10 Analysis of chemical space

The patterns associated with active chemical structures were evaluated by analyzing the compounds which were correctly and incorrectly predicted by respective machine learning models. The target ER-LBD was chosen mainly because the ensemble method achieved the best performance for this target[64]. It is of particularly interesting to investigate which chemical space was correctly or wrongly represented by different molecular fingerprints (ECFP4 and MACCS). All the active chemical structures predicted by the RF model were also correctly predicted by the NB model as illustrated in figure 4.4. Additionally, the NB model predicted five more active compounds correctly whereas the PNN model failed to predict a single active compound. On the other hand, the number of inactive compounds incorrectly predicted (false positives) in NB models was the highest with 80 incorrect predictions, followed by RF with 4. PNN based models predicted all the inactives correctly supporting the fact that it is biased towards majority class coverage as reported in table 4.16. Additionally, it is observed that NB based model with both ECFP4 and MACCS fingerprints predicted the active compounds with highest confidence scores compared to RF models as reported in table 4.13. It could be because RF fails to predict the active class when the molecules become more complex irrespective of the fingerprints considered as shown in figure 4.4.

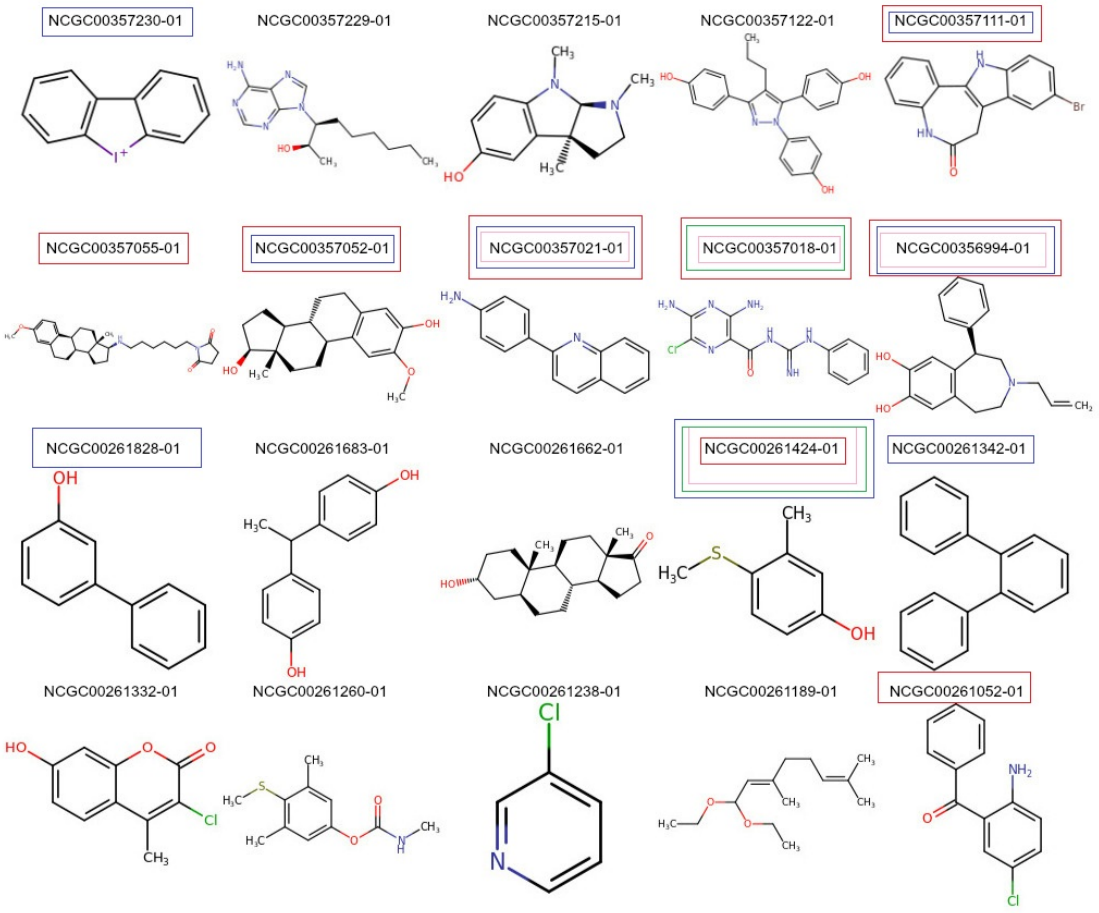


FIGURE 4.4: **Chemical space analysis** The above figure shows the different actives present in the external set of ER-LBD. The compounds highlighted in pink boxes and blue boxes were correctly predicted by Random Forest classifier and Naïve Bayes classifiers respectively. Additionally, respective confidence scores for each classifier are shown.

Molecule id	NB with MACCS	RF with MACCS	NB with ECFP4	RF with ECFP4
NCGC00261424-01	0.99	0.58	1	0.77
NCGC00261052-01	0.57	0.07	0.02	0.12
NCGC00357055-01	0.95	0.01	0.01	0.06
NCGC00357018-01	0.99	0.94	1	0.94
NCGC00357052-01	0.99	0.04	0.99	0.16
NCGC00357021-01	0.99	0.68	0.99	0.31
NCGC00356994-01	0.99	0.52	0.99	0.36
NCGC00357111-01	0.99	0.06	1	0.15
NCGC00261828-01	0.13	0.05	1	0.20
NCGC00261342-01	0.01	0.02	0.99	0.08
NCGC00357230-01	0.04	0.05	0.98	0.02

TABLE 4.13: **ER-LBD active compounds correctly predicted in External set using RF and NB models using MACCS and ECFP4 fingerprints. alone with confidence scores**. Color denotes different molecules illustrated in the figure 4.4

4.11 Comparison with Tox21 challenge top performers

After the Tox21 data challenge winners were announced, a comparison of the top performing models as well as the similarity-based fingerprint method with the machine learning models were carried out. The best prediction achieved in this study is by RF classifier with MACCS fingerprints .

In the Tox21 data challenge, the winning method for the targets NR-AhR and SR-HSE was achieved by Mayr et.al[79]. The models were developed based on deep learning algorithm which comprises of abundance of deep neural network. The descriptors used in this method is a combinations of ECFP fingerprints, physico-chemical descriptors and substructures previously used as toxic alerts.

For ER-LBD, the best method in the Tox21 data challenge was reported by Uesawa Y[80]. The model is based on the ensemble of 200 RF models and around 4,071 dragon as well as MOE descriptors were used for the construction of the model.

When compared the RF model reported in this thesis performed equally good as the the Tox21 top models in terms of AUC values as shown in table 4.14 . When the balanced accuracies were compared, the RF model outperformed all the top models in the Tox21 challenge as shown in the table 4.15. When compared to the first model reported in this study (similarity-based fingerprint), the Random Forest model in the study shows better performance with respect to the AUC and balanced accuracy values.

Method	NR-AhR	ER-LBD	HSE
Similarity-based fingerprint	0.89	0.79	0.85
Random Forest	0.90	0.81	0.86
Tox 21 top models	0.92	0.82	0.86

TABLE 4.14: **External validation result comparison with Tox21 Challenge winners**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis

Method	NR-AhR	ER-LBD	HSE
Similarity-based fingerprint	0.73	0.79	0.80
Random Forest	0.85	0.78	0. 82
Tox 21 top models	0.75	0.55	0.73

TABLE 4.15: **External validation result comparison with Tox21 Challenge winners**. Balanced accuracies of the different models and targets

4.12 Discussion

In this study, two different methods were proposed for the prediction of interference of the Tox21 challenge data set consisting of chemical compounds tested in two major biological pathways; nuclear receptor pathway and stress response pathway. The data was generated in a standard uniform experimental setup which serves as a gold standard data source for evaluating performance of different prediction methods. In the first section, a similarity-based fingerprint method was reported which is based on the *'similarity property principle'* and concatenated fingerprints with molecular descriptors based binary fingerprints. This method performed relatively better in combination of different fingerprints and descriptors. In the second section, machine learning based models were developed in combination of individual fingerprints and as well as in combination with molecular property based descriptors.

It was observed that the machine learning models based on RF classifier achieved best performance when compared to other two machine learning models (NB and PNN) as well as the similar-based fingerprint method. The superior performance of the RF models can be attributed to the different tunning parameters chosen for individual targets (i,e RF algorithm is robust to different tuning parameters). On the other hand, the poor performance of the PNN model can be explained by its strong inclination towards the majority class coverage of the training set. To understand this behavior in detail, the results were analyzed and it is observed that PNN models were able to correctly predict all the true negatives in the external validation with a confidence score higher than 0.9 but failed to correctly predict the actives (true positives) for all the three targets. On the other hand, NB models could predict

the highest number of true positives with confidence scores higher than 0.99 in comparison to the RF and PNN based models. However, NB lacks the prediction ability when it comes to true negatives (inactives /majority class). Taking an example of a result from randomly selected target ER-LBD, this trend was analyzed in details for each models considering two different fingerprints (MACCS; ECFP4) as shown in table 4.16. RF based models are able to identify the patterns associated with respective classes in case of imbalanced data set.

Furthermore, it is toxicity alert encoded fingerprints (ToxPrint) as well as atom state based fingerprints (EState) that failed to obtain consistent performance across most of the targets for various models. This could be due to the fact that the chemotypes in the training set do not match with the pre-defined toxicity alerts encoded in the ToxPrint fingerprints. On the other hand the MACCS fingerprints encodes the similar substructures like the one present in the chemical space of the Tox21 challenge data and therefore performed better and consistent across various machine learning models. This supports the fact that prediction of toxicity can not always be encountered using a global approach (i,e identification of presence of certain toxic alerts in the chemical space). Target specificity and local substructures closely associated with the chemical space addressed in the study play an important role in the prediction process. In addition, selection of optimal descriptors which could represent the chemical space and an unbiased classifier which can learn the patterns and used this knowledge to predict activity of an unknown compound is in real a true essence of predictive science.

Overall, in this study, it can be emphasized that a simple RF based classifier consistently demonstrated robust prediction for all the three targets. This method is relatively simpler and computationally less expensive than other methods of the Tox21 challenge. The results of this study are equally good when compared with the Tox21 challenge top models with respect to AUC values and are better with respect to balance accuracies. This further adds to the usability of the optimal method developed in this study.

ER-LBD	**True positives (out of 20)**	**True negatives (out of 580)**	**Cross validation AUC**	**External validation AUC**
NB with ECFP4	9	500	0.76	0.71
NB with MACCS	8	468	0.73	0.69
RF with ECFP4	2	574	0.82	0.78
RF with MACCS	4	576	0.83	0.81
PNN with ECFP4	0	580	0.77	0.69
PNN with MACCS	0	580	0.78	0.69

s

TABLE 4.16: **Classifications of actives and inactives in external set by different models for ER-LB**.

4.13 Conclusion

In this study, different computational approaches were developed and compared to predict activity of the chemical compounds. The random forest based *in silico* toxicity prediction method is reported as the best method, emphasizing the importance of predictive toxicology as a fast and reliable way to predict the toxic outcome of chemical compounds. Different methods including the similarity-based as well as machine learning models were evaluated using four different types of fingerprints and property based descriptors on Tox21 challenge data. The results from this study suggest that RF based machine learning models can improve the accuracies of prediction for all the three targets. The models developed in this study were consistent with the top performing models in the Tox21 data challenge in terms of AUC and have outperformed in terms of balanced accuracies. The model and results of this study will be a great importance in the field of *in silico* toxicity prediction as well as for food and drug regulatory agencies in development of decision-making tools to further improve their control over potentially toxic chemical compounds. Additionally, it can be concluded from the result of this study that the combination and application of RF algorithm and substructure based molecular fingerprints (MACCS) can be regarded as a very promising prediction tool for evaluation of toxic effects of new chemicals.

4.14 Availability

Both the similarity-based fingerprint method and Machine learning methods are developed using open source platform KNIME. The similarity based method is made available as KNIME work flow to be used by the scientific community (http://tox.charite.de/tox/index.php?site=links) . The machine learning methods will be made available via publication (once accepted).

Chapter 5

A novel method to predict fatty liver drugs using metabolic network based target identification

Hepatocytes exhibits a wide range of functions extending from removal of toxic substances, homeostatic regulations and synthesis of most plasma membrane constituents as well as production of bile and hormones[81,82]. Hepatocytes have higher metabolic activity in human and plays an important role in human metabolism. Alterations in the metabolism of hepatic cells can lead to complicated liver problems like hepatitis, non alcoholic fatty liver disease (NAFLD), cirrhosis and liver cancer, and can be serious threats to public health[81]. NAFLD is considered as the hepatic manifestation of obesity and metabolic syndrome as a result of series of pathological changes, which ranges from reversible fatty liver (Steotosis) to non alcoholic steatohepatitis (NASH)[81].

In this chapter,the molecular targets which plays important role in the metabolic network are considered and further studied to detect drugs that have the potentiality to cause fatty liver syndrome. Two different hypotheses have been postulated, based on that two different computational prediction methods have been developed, ligand-based pharmacophore model and molecular docking based prediction.

5.1 Introduction

In recent years, the most common cause of chronic liver disease in USA is contributed by NAFLD. A recent study shows the need and cost associated with medication related to liver

diseases as well as organ transplantation[83]. Increase rates of obesity diabetes and high cholesterol have results in growing concern for liver diseases[84]. NAFLD and Steatohepatitis are well linked but rare forms of drugs induced liver injury[85]. In addition, fatty liver is often chronic than acute even when drug induced[86]. Even though it is well known that the lipid accumulation in the liver is a starting point of the NAFLD; the underlying mechanism leading to steotosis is still elusive[81]. The adverse outcome of this pathology may be possibly prevented once the molecular mechanisms involved in the metabolism of liver are unraveled. On the other hand, this requires understanding of the coordinated behavior of a very large number of interconnected network of drugs, molecular target as well as off-target, pathways, metabolic network and metabolites.

Recent developments in the field of computational systems biology made it possible to predict the functional effects of systems perturbations using large scale network models. Subsequently, advances in the field of structural bioinformatics and chem-informatics have led to the prediction of protein-drug off-target effects based on their ligand structures and binding site information. Integration of these expertise provides a platform for evaluating metabolic drug response *in silico*. The combination of these approaches was applied to investigate the drugs that can cause fatty liver disease in human. Currently, there is no efficient treatment or explanation involved in the mechanism of NAFLD.

This study represents a novel integration of the trio 'computational systems biology, structural bioinformatics and cheminformatics' approaches to predict drugs involved in metabolic syndrome and to understand the possible underlying mechanism.

5.2 Metabolic Network

The primary goal of cellular metabolism is the conversion of small molecules by intracellular catalysts (enzymes) to provide life maintaining support like growth and reproduction[87]. Enzymes can speed up certain biochemical reactions such as conversion of substrates into product often by attaching chemical group or breaking off chemical group from the substrates. Metabolic network can be termed as a simplified representation of this complex network of biochemical reactions. The aim of the metabolic network is to unravel the basic principles, predict cellular responses or to identify important metabolic targets for intervention[88,89].

Kinetic modeling is special kind of metabolic network representation[90,91]. Kinetic models represents metabolic network by a stoichiometric matrix S_{ij} connecting metabolites M_i and reactions v_j. The time dependent changes in metabolite concentrations $[M_i]$ are given by

$$\frac{d[M_i]}{dt} = \sum_{i=1}^{n} S_{ij} v_j$$

The reaction rates v_j are dependent on external and internal parameters. Regulatory principles considers compromise kinetic regulation through substrate availability as well as allosteric regulation. Additionally changes in enzymes structures or in abundance is also taken into account. In a given cellular compartment, local enzyme and metabolite concentrations are replaced by mean concentrations and it is assumed that stochastic variations can be ignored. Given the rate of equations for the v_j are constant, only additional parameters like external medium compositions such as metabolite, ion or hormone concentration are required for the model.

Such model can be applied to calculate the concentration of metabolites and the rates of their mutual chemical interconversion in response to varying external conditions like nutrient supply and varying internal conditions like enzymes availability and demand. These calculations are further utilized to quantify complex cellular functions as cellular growths, detoxification of drugs and xenobiotic compounds or synthesis of exported molecules[87].

The kinetic model used in this study includes the main pathways of carbohydrate, amino acid and fatty acid metabolism comprising: glycolysis, fructose metabolism, galactose metabolism, gluconeogenesis, glycogen metabolism, citric acid cycle, fatty acid synthesis, cholesterol synthesis, beta-oxidation, respiratory chain, oxidative phosphorylation, mitochondrial electrophysiology, malate-aspartate shuttle, glycerol-3-phosphate shuttle, urea cycle, glutamine synthesis, glutaminolysis, serine metabolism, alanine metabolism, triacylglycerol synthesis, lipid storage and mobilization, VLDL assembly and secretion, as well as ketone body production. The model describes the exchange of the plasma metabolites glucose, galactose, fructose, pyruvate, lactate, glycerol, acetate, beta-hydroxybuterate, free fatty acids, ethanol, acetoacetate, oxygen, glutamine, glutamate, serine, alanine, ammonia as well as the storage of glycogen and triglycerides. The metabolic system takes kinetic effects on inter convertible enzymes in response to the hormones insulin and glycogen into account.

5.3 Selection of target

The above kinetic model based on the central hepatic metabolism was used to identify enzymes which upon inhibition lead to an increased fat accumulation within the hepatocytes. The systematic inhibition of each enzymes by 10 % and monitoring of the resulting triglyceride content in a healthy hepatocyte under physiological conditions revealed the top ranked enzymes and the corresponding pathways as shown in the table 5.1 .

Enzyme	Functioning pathways
apoB synthesis	VLDL
Microsomal transfer protein	Lipid droplet synthesis
Glucose-6-phosphate phosphatase	Gluconeogenesis
Fructosebisphosphatase 2	Gluconeogenesis
Carnitine palmitoyl transferase I	Fatty acid oxidation
Pyruvate kinase	Glycolysis
Malonyl-CoA decarboxylase	Fatty acid synthesis

TABLE 5.1: **Top 10 ranked enzymes from the kinetic model of central hepatic metabolism**

.

Several literature sources were search using a datamining workflow based on KNIME; to obtain information on the important targets from the metabolic network. It was found most of the targets had no crystal structures (human) as well as known inhibitors. It was found that inhibition of acetyl-Coa carboxylase (ACC) results in inhibition of fatty acid synthesis and stimulation of fatty acid oxidation. Therefore, it has the potential to favorably effect metabolic syndrome[92,93].

On the other hand, carnitine palmitoyl transferase (CPT1) plays an important role in fatty acid oxidation[94], including a rate limiting role by its inhibition by malonyl-CoA[95]. Since CPT1 inhibitors were publicly available with strong affinities values, it was selected as the first target in the network to investigate further.

5.4 Fatty acid oxidation

The enzymes of involved in fatty acid (FA) oxidation are located in the mitochondrial matrix in animal cells, as demonstrated by Eugene P.Kennedy and Albert Lehninger[96]. The small chain fatty acids (12 or fewer carbons) can enter the mitochondrial membranes without membrane transporters. Wheres as, the long chain fatty acids (14 or more carbons) which constitute the majority of the FA obtained in the diet or released from adipose tissue, cannot directly enter the mitochondria membranes and hence, they must undergo the three enzymatic reactions of the carnitine shuttle[96,97].

The first reaction is catalyzed by a family of isoenzymes present in the outer mitochondrial membrane, the acyl-CoA snythetases, which promotes the reaction. Thus acyl-CoA synthetase catalyze the formation of a thioester linkage between the fatty acid carboxyl group and the thiol group of coenzyme A to yield fatty acyl-CoA[96]. In the second step fatty acyl-CoA esters formed at the cytosolic side of the outer mitochondrial membrane are transported to the inner mitochondrial membrane and oxidized to produce ATP. Fatty acids destined for mitochondrial oxidation are transiently attached to the hydroxyl group of carnitine to form fatty

acyl-carnitine[96] catalyzed by carnitine acyltransferase I. The fatty acyl-carnitine ester enters the matrix through the acyl-carnitine transporter of the inner mitochondrial membrane.

In the final step, the fatty acyl group is enzymatically transfered from carnitine to intra-mitochondrial coenzyme A by carnitine acyltransferase II[96,97]. This isoenzyme located on the inner face of the inner mitochondrial membrane, regenerates fatty acyl-CoA and releases it, along with free carnitine into the matrix. Carnitine re-enters the membrane space via the acyl-carnitine/carnitine transporter[96].

This carnitine-mediated entry process is the rate limiting step for oxidation for fatty acids in mitochondria and plays a regulating role[96].

On the other hand, acetyl-coenzyme A carboxylase 1 (ACC1) and acetyl-coenzyme A carboxylase II (ACC2) plays important role in lipid metabolism. Malonyl- CoA is produced due to catalysis of the acetyl-CoA by ACC1 in the cytosol and acetyl CoA is utilised via fatty acid synthase (FAS) reactions to generated palmitate, and further utilized in the synthesis of triglycerides (TG) and VLDL. Additionally, acetyl-CoA is carboxylated by ACC2 at the mitochondrial membrane to form Malonyl CoA; which inhibits the acyltransferases I[93]. This inhibition prevents the simultaneous synthesis and degradation of fatty acids[96].

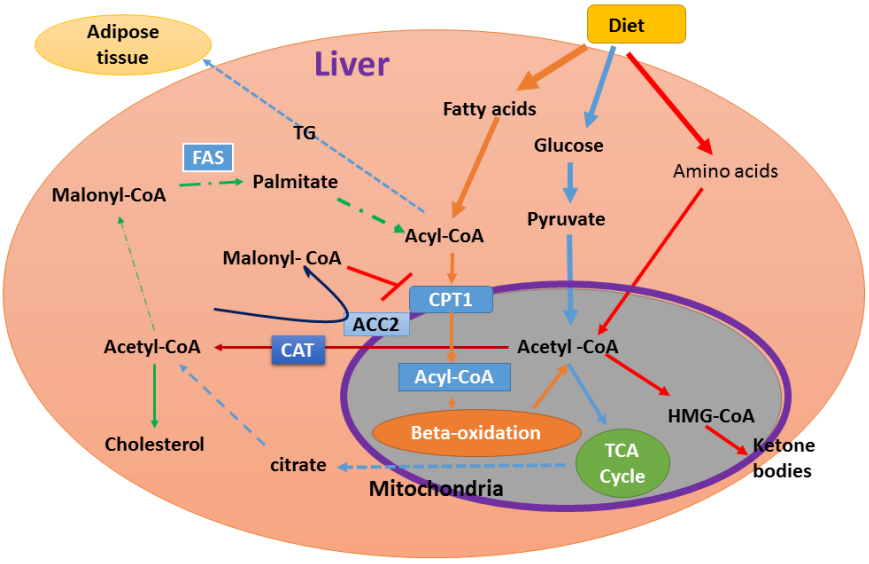


FIGURE 5.1: **A schematic representation of the process of fatty acid oxidation** The above figure shows different enzymes involved in the fatty acid oxidation process and their respective influences in the process.

5.5 Formulation of hypothesis

There are two hypotheses that are postulated in this study and are believed to play important role in the mechanism of drug induced fatty liver disease. First hypothesis is based on the

malonyl CoA binding site of the CPT1 A and the second one is based on the carnitine binding site.

5.5.1 First hypothesis

The first hypothesis is related to the inhibitory activity of malonyl-CoA[97]. This is well established fact that inhibition of the CPT1 enzyme by malonyl-CoA is key limiting factor in the process of fatty acid oxidation. This information is used to select drugs which have similar pharmocophoric properties like malonyl-CoA and therefore could bind to the allosteric site of the CPT1 and results in its inhibition[98].

Thought it is completely not clear how the mechanism of inhibition of CPT1 takes place by malonyl-CoA; however many studies have suggested that malonyl binds either through 'A site' or 'O site' of the CPT1 enzyme[97]. Therefore, it might interact with the carnitine binding site of the CPT1 which is important for shuttling of the acetylated fatty acids across other mitochondrial membrane for oxidation inside the mitochondria[98].

5.5.2 Second hypothesis

CPT1 in its liver isoform has higher affinity for carnitine and it is found that the carnitine based active site inhibitors are highly studied in literature[95]. The unavailability of the crystal structure of CPT1A isoform makes it difficult to developed the structural based model of the enzymes and further validate with computational docking studies. However, the class of carnitine acyltransferases enzymes which includes acyltransferase (CrAT), Octanoyltransferase (CrOT) and carnitine palmitoyltransferases (CPTs) are highly homologous. It been suggested in many studies that the catalytic domains of these enzymes are well conserved, with 35 % or higher amino acid sequence identity between any pair of them[99].

The carnitine binding site of the CrAT protein is located at one of the entrance of the active site tunnel[100]. The key residues that plays important role in the molecular binding interactions are His322, Tyr 431, Thr 444, Arg 497, Thr 444, Arg 497, Phe 545, Val 548, Ser 433 and two tightly associated water molecules ($H20^{621}$ and $H20^{679}$)[100]. Among these active sites residues, His322 is completely conserved in all the carnitine acyltransferases[100]. It has been reported that carnitine binds to the active site of the protein with its beta-hydroxyl moiety aligned directly with the His322 residue to form a hydrogen bond with the N-3 atom of the imadazole ring[100]. This interactions is important for the catalytic role of the His322, which has been proposed to deprotonate the beta-hydroxyl group of the carnitine or the thiol group of the CoA to facilitate nucleophilic attack on the ester linkage of the acyl group[100]. Thr 444 interacts with one of the oxygen of the carboxylate group[100]. Additionally, other important

residues in the active sites are Trp81, Tyr86, Tyr431 and Glu 326 binds to carnitine with forming hydrogen bonds[100]. It has been reported that hydrogen bonding interaction seems to be a primary force for the recognition and binding of L-carnitine[100]. Moreover, the positive charge in the carnitine is not required for binding, but it is required for catalysis[100]. This positive charged N atom is important for catalysing the transfer of the acyl group of a long-chain fatty acyl-CoA from CoA to I-carnitine[100].

Additionally, the alignment of the L-carnitine and His322 is stereo specific and this explains why D-carnitine is the inactive enantiomer, since the beta-hydroxyl group of the D-carnitine would point in the opposite direction of Hiss 322 and therefore, would not be accessible for acyl group transfer[100].

The main assumption here is drugs that are similar to carnitine and have required functional group for binding may bind in the carnitine binding site of the CPT1. However, lack of binding which favored the catalytic role of the enzyme by taking into the most important residues as well as absence of the positively charged N atom, it will not be able to continue catalytic process and results in no fatty acid transfer for beta-oxidation in the mitochondria and thus, leading to accumulation of fatty acids outside. This will result in competitive inhibition of CPT1. One argument in support of this hypothesis can be stated that, Oxfecine and Pisolithin A (pubchem id :36143; 355) which are structurally very similar to carnitine with an additional 6-membered ring in it; is also an inhibitor of CPT1. On the hand, Emeriamine also known Amino-carnitine/ Emeriamine (pubchem id:121830) is also an inhibitor of CPT1 as shown in figure

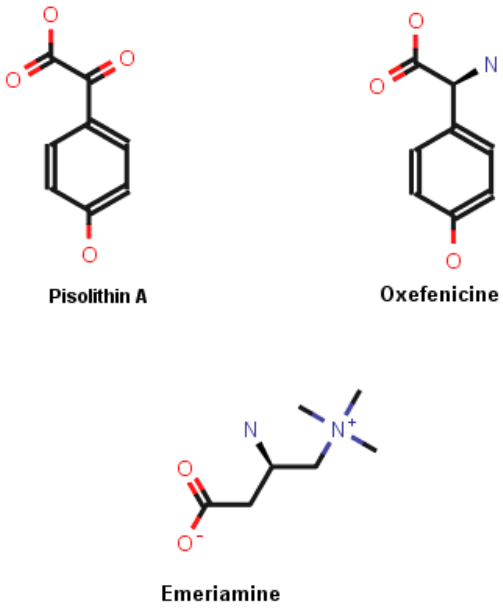


FIGURE 5.2: **Structure of inhibitors of CPT1 similar to carnitine**.

5.6 Data set processing

5.6.1 Pharmacophore model for drugs similar to malonyl CoA

The ligand based pharmacophore model is highly dependent on an accurate and precise selection of training set. In course of data selection it was considered that most, medium and least active compounds were incorporated in the training set and test as shown in the tables 5.2 and 5.3. A total of 15 training set compounds were selected. The compounds with activity IC_{50} value less than 1 mM were considered as most active and rest as medium actives.

7 test compounds were selected for the validation of the pharmacophore model. In the test compound sets the compounds similar to carnitine (CHEMBL2114397, CHEMBL129918, CHEMBL75880, CHEMBL1232077, CHEMBL203266) were considered as inactives. This was done because, the pharmacomodel is based on the malonyl Co-A binding site and so molecules with higher molecular weight were considered as most actives (This is also confirmed by the IC $_{50}$ values. However, there were two active compounds in the test set (CHEMBL633, CHEMBL2216773). One of the active compound in the test is amiodarone (CHEMBL633) with an IC_{50} value of 140 mM. Amiodarone is considered as one of the classical fatty liver drugs [101].

Number	CHEMBL id	Pubchem id	activity IC_{50} in mM	Pubmed id
1	CHEMBL2216778	16718554	0.016	21504156
2	CHEMBL2216774	25095157	0.02	21504156
3	CHEMBL2216779	15984085	0.02	21504156
4	CHEMBL2216776	66857424	0.026	21504156
5	CHEMBL2216775	16736346	0.065	21504156
6	CHEMBL2216777	124825719	0.165	21504156
7	CHEMBL2216780	11979264	0.19	21504156
8	CHEMBL2216769	71457921	0.24	21504156
9	CHEMBL2216768	23695524	0.25	21504156
10	CHEMBL2216782	71452605	1	21504156
11	CHEMBL2216781	71463291	1	21504156
12	CHEMBL1231506	9843897	1.5	21504156
13	CHEMBL2216785	23729161	1.7	21504156
14	CHEMBL2216771	15234003	2.4	21504156
15	CHEMBL1231507	13671155	2.67	21504156

TABLE 5.2: **Training set for pharmacophore model**. The table shows the total number of training set compounds and their activity value

The training set is constructed using information from published source [95]. All the inhibitors were experimented on a same essays and under similar experimental conditions. The chemical structures of the all the 15 training set compounds used for generation of the model is shown in the figure 5.3 .

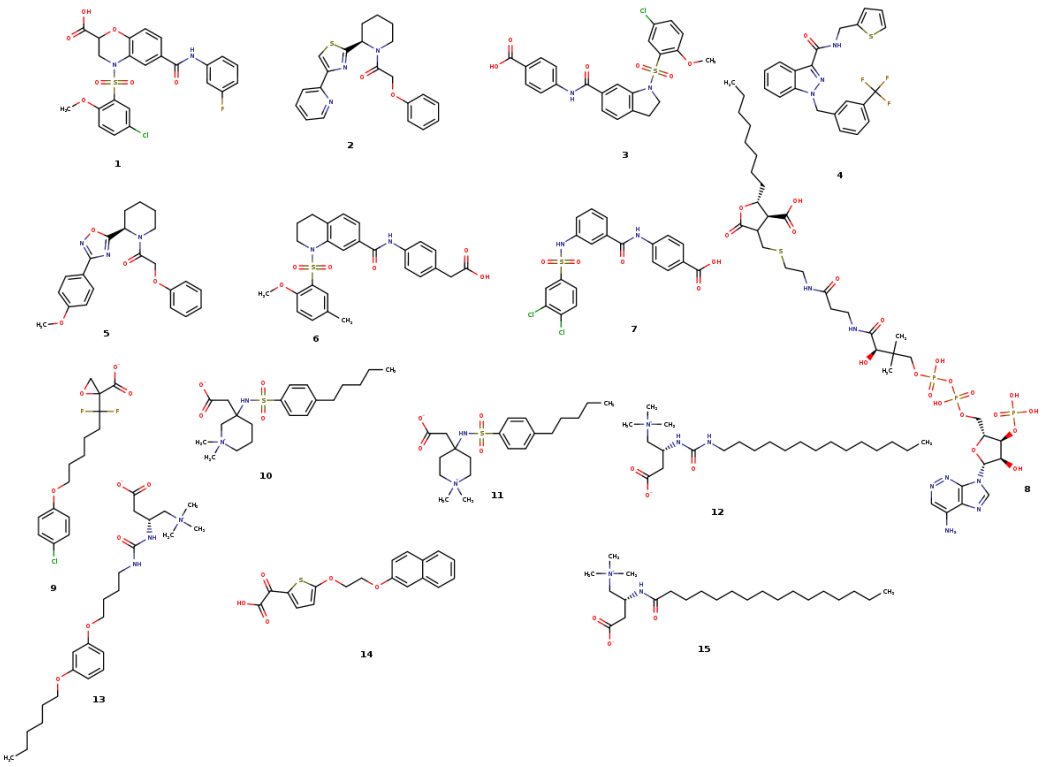


FIGURE 5.3: **Structure of ligands in the training set** Most of the active compounds are large molecules and has similar functional groups

Number	CHEMBL id	Pubchem id	activity in mM	Pubmed id
1	CHEMBL2114397	121830	76.4	21504156
2	CHEMBL129918	355	>100	21504156
3	CHEMBL75880	4746	>100	21504156
4	CHEMBL1232077	36143	>100	21504156
5	CHEMBL203266	21109	100	21504156
6	CHEMBL633	2157	140	21504156
7	CHEMBL2216773	9893640	51000	21504156

TABLE 5.3: **Test set for pharmacophore model**. The table shows the total number of test set compounds and their activity value

The test set compounds were also extracted from the same published literature source and were experimented in same essays[95]. The chemical structures of the all the 7 test set compounds used for generation of the model is shown in the figure 5.4.

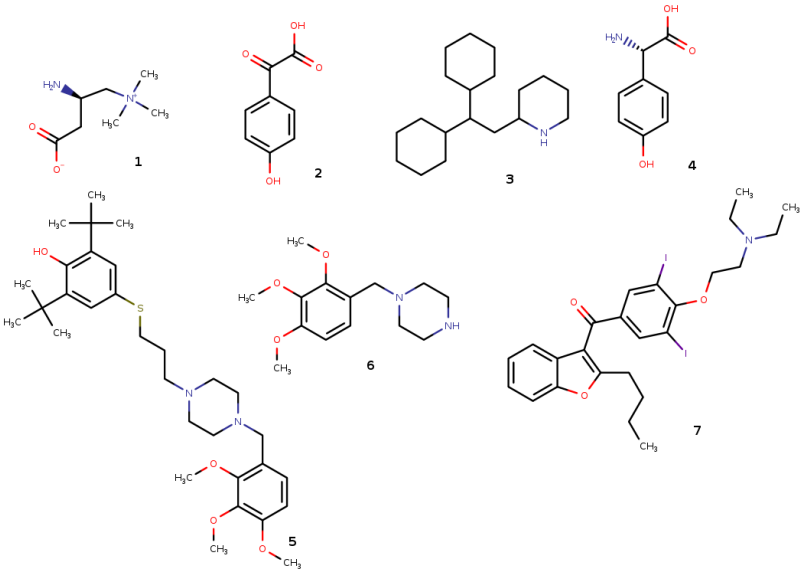


FIGURE 5.4: **Structures of the test set**. The compounds similar to carnitine and smaller in size are considered as inactive

5.6.2 Computational docking of drugs similar to carnitine

The data set for this part was extracted from drugbank database. Since the lead compound was carnitine and the screening dataset includes all approved drugs.

5.6.3 Known fatty liver drug set

Based on data mining and text mining from literature sources with queries like fatty liver, steotosis, steotohepatitis, liver warning, hepatic warning, drug-induced fatty liver; a total of 12 drugs associated with fatty liver disease were retrieved.

5.7 Methods

In this section two methods for each of the hypotheses have been explained in details.The pharmacophore based model for predicting drugs that could bind to the malonyl-CoA binding site of the CPT1 and a computational docking based model for predicting drugs that could bind to the carnitine binding sites of the CPT1.

5.7.1 Ligand based pharmacophore modeling of malonyl CoA similar inhibitors

Ligand based pharmacophore model represents a key computational strategy in the absence of a macromolecular target structure. A common feature pharmacophore model was developed using information from a set of active compounds for the target CPT1. The model is heavily dependent on the extraction of the common chemical features from 3D structures from the set of known ligands which represents essential interactions between the ligands and the molecular target CPT1. The algorithm behind this model is explained in details in the section 2.4.

Diverse conformation generation protocol was executed using Discovery Studio software. The parameters for number of conformers were kept as 300 conformations and an energy threshold of 29kcal/mol above the global energy minimum were considered f or generating the conformers. Prior to hypothesis generation, common pharmacophore features amongst the active training set compounds were analyzed using automatic pharmacophore generation protocol in Discovery studio as shown in figure 5.5 .

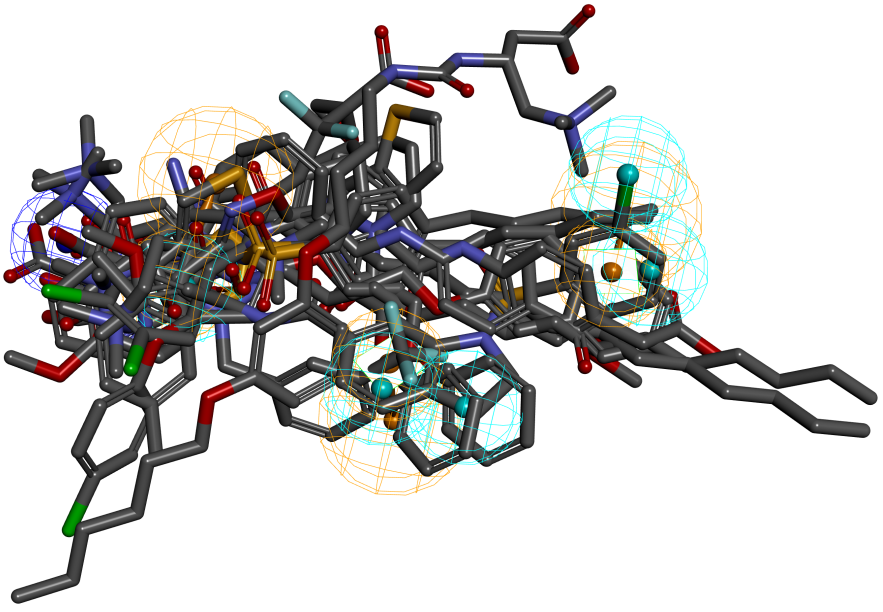


FIGURE 5.5: **Common pharmacophoric features present in the training set** The negative ionisable feauture is shown in blue, hhydrophobic feature is cyan and ring aromativ feature in orange

The qualitative top ten hypotheses were generated based on the training set active molecules using the *Common Features Pharmacophore Generation* in Discovery Studio. The common feature were identified which are necessary to inhibit CPT1. The direct hit value of '1' and partial hit value '0' indicates that the molecules present in dataset are well mapped to all the chemical features in the hypothesis and there is no partial partial mapping or missing features as specified in the table 5.4. The external validation result are shown in the table

5.5. According to the external validation results all the four hypotheses (1,2,3,4) were able to predict the true actives with sensitivity 1 and specificity 0.80. The best hypothesis (Hypo1) and its pharmacophore features were chosen to predict new compounds that could inhibit CPT1.

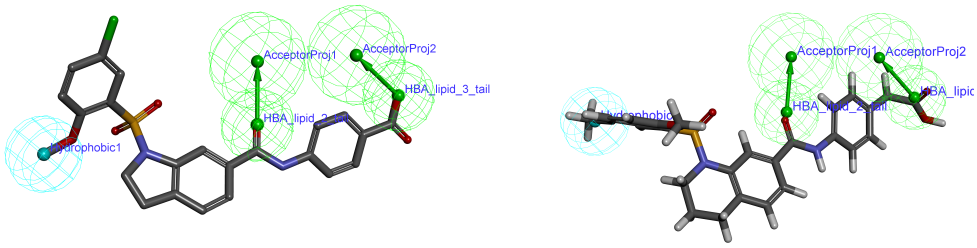
Model	Features	Rank	Direct Hit	Partial Hit	Max Fit
Hypo1	ZHH	118.97	111111111111111	000000000000000	3
Hypo2	ZHH	118.24	111111111111111	000000000000000	3
Hypo3	ZHH	117.97	111111111111111	000000000000000	3
Hypo4	ZHH	117.57	111111111111111	000000000000000	3
Hypo5	ZHA	116.17	011111111111111	100000000000000	3
Hypo6	ZHA	116.17	011111111111111	100000000000000	3
Hypo7	ZHA	115.44	011111111111111	100000000000000	3
Hypo8	ZHA	115.44	011111111111111	100000000000000	3
Hypo9	ZHA	115.17	011111111111111	100000000000000	3
Hypo10	ZHA	115.17	011111111111111	100000000000000	3

TABLE 5.4: **The result of top 10 hypotheses generated by HipHop program**. Three features were found to be most common on the active compounds in the training set, one feature corresponds to 'hydrophobic' (Z) and two features corresponding to 'hydrogen bond acceptor lipid' (H). The higher ranking scores indicates the lesser the probability of chance correlation. The best hypothesis is indicated with the highest value

Pharmacophore	number of actives	number of inactives	Sensitivity	Specificity
Hypo1	2	5	1	0.80
Hypo2	2	5	1	0.80
Hypo3	2	5	1	0.80
Hypo4	2	5	1	0.80
Hypo5	2	5	0.50	0.60
Hypo6	2	5	0.50	0.60
Hypo7	2	5	0.50	0.60
Hypo8	2	5	0.50	0.60
Hypo9	2	5	0.50	0.60
Hypo10	2	5	0.50	0.60

TABLE 5.5: **Validation of the all the ten hypotheses on an external validation set**.

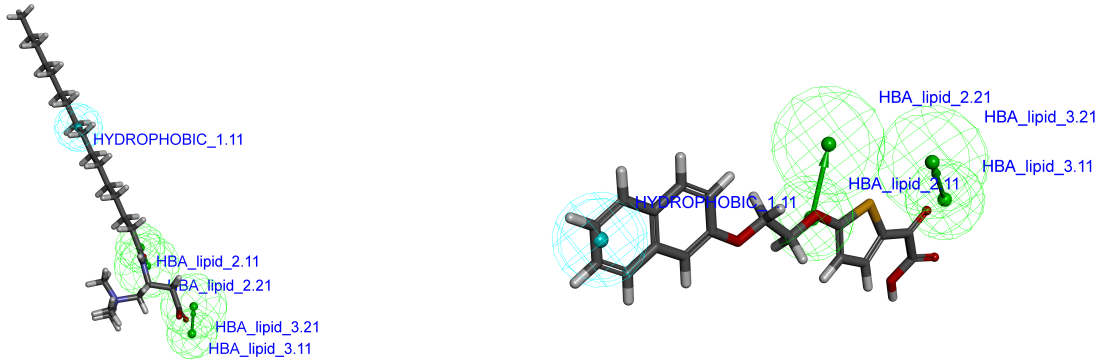
To understand how the actives and least actives were mapped with the pharmacophore features generated by Hypo1, all the training set compounds were visualized individually and their mapping were noted. Some of the most actives training compound and least active training compounds mapped with the pharmacophore features of Hypo1 are shown in the figures



(A) Pharmacophores of molecule '16718554'

(B) Pharmacophores of molecule '24825719'

FIGURE 5.6: **Pharmacophore alignment with the active training set molecules**



(A) Pharmacophores of molecule '13671155'

(B) Pharmacophores of molecule'15234003'

FIGURE 5.7: **Pharmacophore alignment with the least active training set molecules**

Furthermore, the small set of already known fatty liver which was extracted from literature sources using text mining, were mapped with the pharmacophoric features of the hypo1. Out of 12 compounds in the fatty liver drug set, 4 compounds were able to mapped as shown in the figure 5.8 .

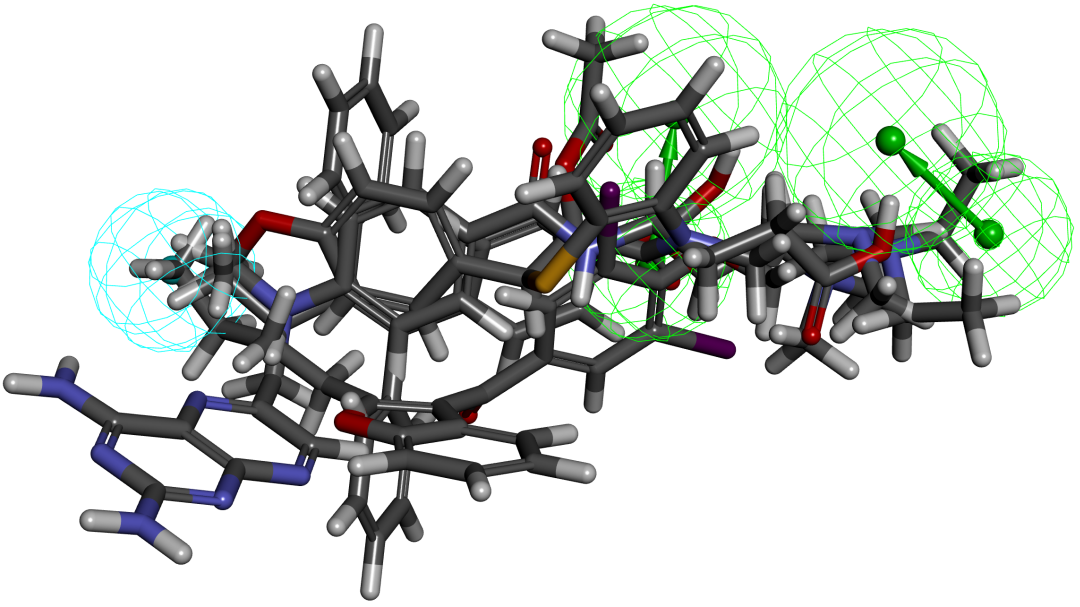


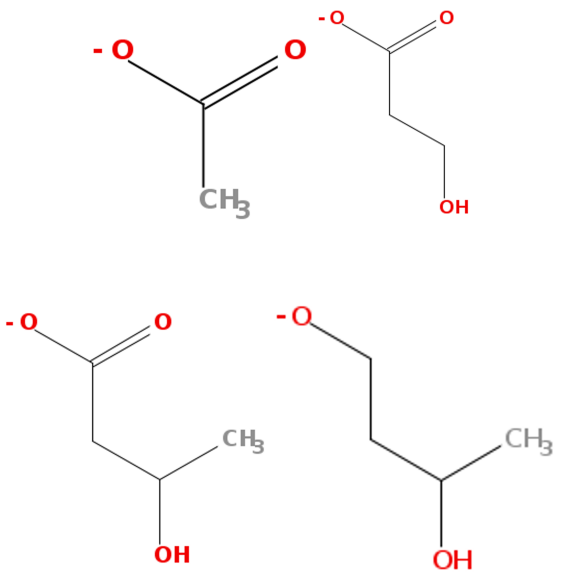
FIGURE 5.8: **Common pharmacophoric features of Hypo1 to the known fatty liver drugs**, The hydrophobic feature is indicated in cyan and hydrogen bond acceptor lipid shown in green. Fatty liver drugs represented in this set are amiodarone, methrotrexate, tamoxifen and diltiazem

5.7.2 Computational molecular docking based on carnitine similar inhibitors

The structural homology model of CPT1 has been developed based on the mouse CrAT crystal structure[102]. Since the crystal structure of human CrAT protein was available (pdb ID:1S5O), this structure was used for the computational docking studies in this thesis. The assumption in this study is since the carnitine binding site of the class of carnitine acyltransferases enzymes are highly conserved, so drugs binding to carnitine binding site of the CrAT protein may also bind to the carnitine binding site of the CPT1.

In order to screen drugs similar to carnitine, two different search methods were implemented. It is previously known that hydroxyl group and both the oxygen of the carboxylate group of carnitine are important for binding. This information was used for a local search criteria. Carnitine molecule was fragmented into 54 fragments with minimum fragment length of 3 atoms to maximum fragment length of 7 atoms. The fragments were further reduced to four by considering only those patterns (functional groups) which were important for binding. These four fragments were used for substructure search[103] against the set of approved drugs. The substructure search were computed using exact atom charges and stereo match. Each of the drug containing any one of the fragments were given a score of 1.

Carnitine fragments



A global similarity score was calculated between carnitine and approved drugs using MACCS molecular fingerprint based similarity search. Tanimoto score was considered as the measure of global similarity with a threshold of 0.50 and above.

Name	Pubchem id	FragScore	GlobalScore
Hydroxybutanoate	3037032	1	0.70
Cefepime	5479537	1	0.60
Ceftazidime	5481173	1	0.50
Methacholine	1993	1	0.84
Bethanechol	2370	1	0.63
Valproate	3121	1	0.68
Bromfenac	60726	1	0.55
Ibuprofen	3672	1	0.56

TABLE 5.6: **Screened drugs based on fragment and similarity search scoring**.

Molecular docking of the screened drugs into the binding site of CrAT were performed using Gold docking software and GoldScore scoring function. The active site cavity was set to 6 Å).The docking results were further visually analyzed using PyMol software and ligand interactions plots were obtained using Discovery Studio software.

5.8 Results

The results obtained from two different methods based on different hypotheses are reported in the individual sections below.

5.8.1 Results of pharmacophore model

The Hypo1 which was ranked as the best hypothesis was considered to as query for predicting new drugs that could inhibit CPT1. All the approved drugs set from DrugBank were considered as a dataset. The multiple conformers of this dataset were computed using best confirmation generation protocol in the discovery studio. The pharmocophoric features of the Hypo1 were mapped with the approved drug set. The resulting molecules were ranked according to their geometric fit value that indicates how well the molecules were mapped onto the hypothesis features location constraints and their distance deviation from the feature centers. The compounds with the highest fit values are suggested for experimental validation.

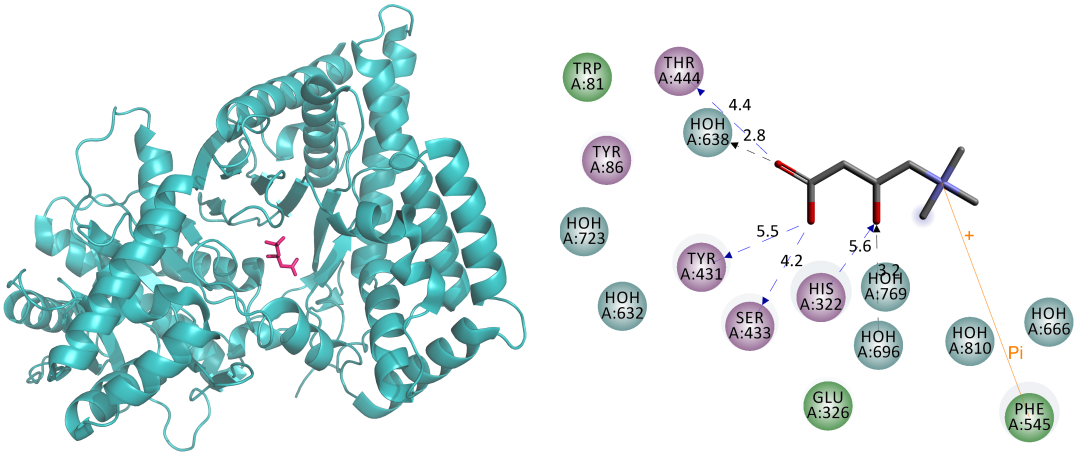
Out of 2399 FDA approved drugs in drug bank database, 100 drugs were mapped with the pharmacophore features with best fit value of 2. 86 and above. The max fit value achieved by the screen drugs was 2.99. The top 30 drugs were further shortlisted based on the best fit value as shown in the table 5.7 .

Number	Name	Pubmed id	Fit value	Therapeutic area
1	Nafarelin	25077649	2.991	Gonadotropin-releasing hormones
2	Carfilzomib	11556711	2.991	Antineoplastic agents
3	Fexofenadine	3348	2.990	Respiratory system
4	Ibutilide	607536	2.990	Antiarrhythmics
5	Epoprostenol	5280427	2.989	Antithrombotic agents
6	Ritonavir	392622	2.989	HIV protease inhibitor
7	Treprostinil	6918140	2.988	Antithrombotic agents
8	Atazanavir	148192	2.980	HIV protease inhibitor
9	Raltitrexed	104758	2.980	Antineoplastic agents
10	Azilsartan medoxomil	11238823	2.971	Cardiovascular system
11	Dirigestran	36523	2.963	Gonadotropin-releasing hormones
12	Xarator	4636594	2.960	Cardiovascular system
13	Formoterol	3410	2.947	Respiratory system
14	Haloperidol	3559	2.947	Anti-psychotic
15	Eltrombopag	9846180	2.931	Antihemorrhagics
16	Dabigatran Etexilate	6445226	2.93	Antithrombotic agents
17	Montelukast	5281040	2.914	Leukotriene receptor antagonists
18	Bosentan	104865	2.907	Antihypertensives
19	Terlipressin	72081	2.895	Vasopressin and analogues
20	Desmopressin	16051933	2.895	Vasopressin and analogues
21	Ximelagatran	18670936	2.890	Antithrombotic agents
22	Zafirlukast	5717	2.890	Leukotriene receptor antagonists
23	Tipranavir	54682461	2.890	Protease inhibitors
24	Tamsulosin	129211	2.890	Alpha-adrenoreceptor antagonists
245	Amiodarone	2157	2.890	Antiarrhythmics
26	Salmeterol	5152	2.890	Respiratory system
27	Cromolyn	2882	2.890	Anti-inflammatory
28	Benziodarone	6237	2.889	Cardiovascular system
29	Glimepiride	3476	2.889	antidiabetic
30	Methotrexate	126941	2.889	Antineoplastic agents

TABLE 5.7: **Top 30 screened drugs with highest fit values predicted by model based on Hypo1**.

5.8.2 Results of computational docking

Based on the molecular docking experiments final three drugs are suggested that can bind to the carnitine binding site of CPT1. Upon binding of these drugs, carnitine dependent transfer of the fatty acid into the mitochondria will be affected and hence result in fats accumulation.

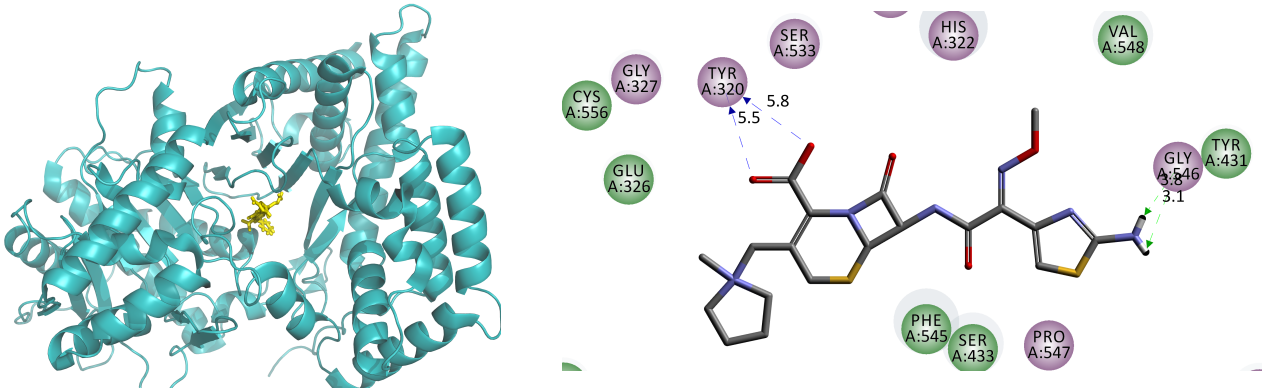


(A) Carnitine binding to the CrAT protein

(B) Ligand interaction diagram of carnitine and CrAT

FIGURE 5.9: **Carnitine binding interaction with CrAT protein**

Cefemine (yellow) as shown in the figure 5.10, binds to the active site of CrAT (blue). The ligand interaction analysis of cefepime with CrAT protein shows that the oxygen of the carboxylate group of cefepime interacts with Tyr320 residue of the protein forming hydrogen bond interactions. Similarly, the terminal nitrogen of the cefepime interacts with Gly546 and Tyr421 residues of the protein.

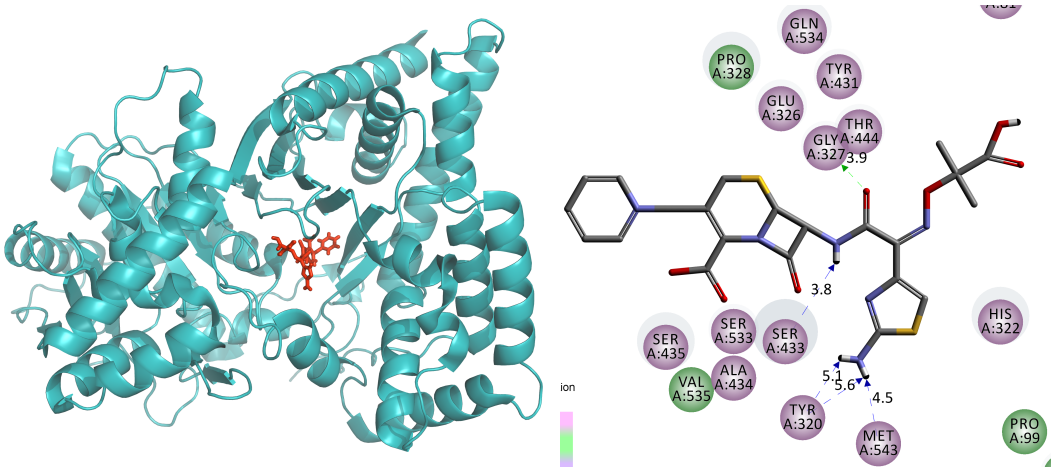


(A) Cefepime binding to the CrAT protein

(B) Ligand interaction diagram of cefepime and CrAT

FIGURE 5.10: **Cefepime binding interaction with CrAT protein**

Ceftazidime (red) as shown in the figure 5.11, binds to the active site of the CrAT (blue). From the ligand interaction diagram it can be seen that nitrogen of ceftazidime interacts with Ser433 residue of the protein forming hydrogen bond interaction. Additionally, the terminal nitrogen of ceftazidime interacts with Met543 and Tyr320 of the protein.

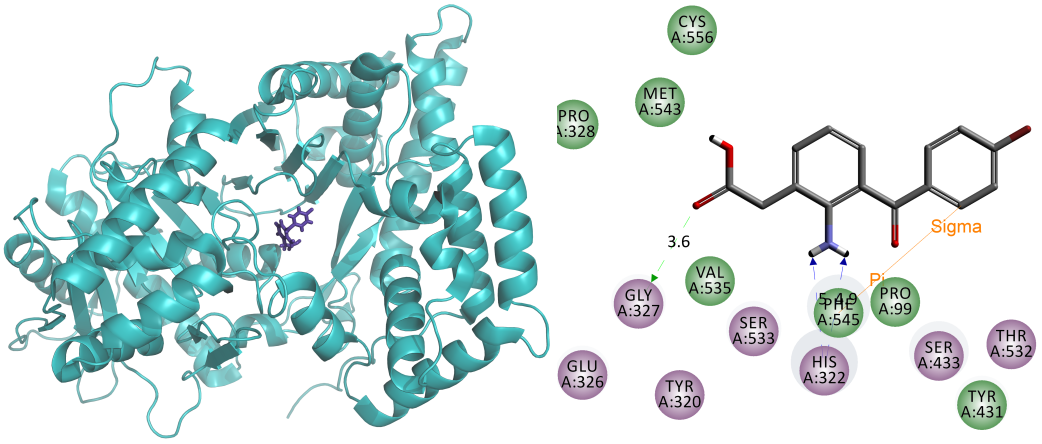


(A) Ceftazidime binding to the CrAT protein

(B) Ligand interaction diagram of Ceftazidime and CrAT

FIGURE 5.11: **Ceftazidime binding interaction with CrAT protein**

Bromfenac (purple) as shown in the figure 5.12, binds to the active site of the CrAT protein (blue). The ligand interaction plot shows the nitrogen of the bromfenac interacts with His322 residue of the protein forming hydrogen bond interaction. The terminal ring of the bromfenac forms pi-sigma interaction with Phe545 residue of the protein.



(A) Bromfenac binding to the CrAT protein

(B) Ligand interaction diagram of bromfenac and CrAT

FIGURE 5.12: **Bromfenac binding interaction with CrAT protein**

5.9 Predicted fatty liver drugs

According to the liver tox database of NIH (http://livertox.nih.gov/Phenotypes_Fatty.html), most of the the medications commonly implicated in causing fatty liver includes antidepressant, cortico steroids, anti-psychotic, antiarrhythmic, antineoplastic drugs such as amiadorone, tamoxifen, methotrexate[101]. One of the reason for side-effects of a drug in the non-selective binding to other target besides the intended therapeutic target. Human

genome project revealed biggest number of large paralog families among all sequenced organisms. Since, the paralogous proteins have similar structure, a drug that targets one paralog is likely to bind to other paralogs too. However, the affinity towards these off-targets (paralogs) can be lower than to the intended therapeutic molecular target ; but with the increase of dose and duration of the administration of the drugs , affinity to off-targets will increase and can be high enough to mediate side effects. So, based on the therapeutic index area and daily dose of administration of the drugs (http://www.whocc.no/atc_ddd_index/), the final list of 23 drugs predicted using pharmacophore model are planned for experimental validation as shown in the table 5.8 .

Number	Name	Pubmed id	Fit value	Therapeutic area
1	Carfilzomib	11556711	2.991	Antineoplastic agents
2	Fexofenadine	3348	2.990	Respiratory system
3	Ibutilide	607536	2.990	Antiarrhythmics
4	Epoprostenol	5280427	2.989	Antithrombotic agents
5	Ritonavir	392622	2.989	HIV protease inhibitor
6	Treprostinil	6918140	2.988	Antithrombotic agents
7	Raltitrexed	104758	2.980	Antineoplastic agents
8	Dirigestran	36523	2.963	Gonadotropin-releasing hormones
9	Xarator	4636594	2.960	Cardiovascular system
10	Formoterol	3410	2.947	Respiratory system
11	Haloperidol	3559	2.947	Anti-psychotic
12	Eltrombopag	9846180	2.931	Antihemorrhagics
13	Dabigatran Etexilate	6445226	2.93	Antithrombotic agents
14	Montelukast	5281040	2.914	Leukotriene receptor antagonists
15	Bosentan	104865	2.907	Antihypertensives
16	Terlipressin	72081	2.895	Vasopressin and analogues
17	Ximelagatran	18670936	2.890	Antithrombotic agents
18	Zafirlukast	5717	2.890	Leukotriene receptor antagonists
19	Tipranavir	54682461	2.890	Protease inhibitors
20	Tamsulosin	129211	2.890	Alpha-adrenoreceptor antagonists
21	Cromolyn	2882	2.890	Anti-inflammatory
22	Benziodarone	6237	2.889	Cardiovascular system
23	Glimepiride	3476	2.889	antidiabetic

TABLE 5.8: **Top 30 screened drugs with highest fit values predicted by model based on Hypo1**.

Based on the integration of fragments score (local), similarity score (global) and molecular

docking studies, three new drugs are suggested that could bind to the carnitine binding site of the CPT1 as shown in the table 5.9 .

Name	Pubchem id	FragScore	GlobalScore	Therapeutic area
Cefepime	5479537	1	0.60	Antibacterial
Ceftazidime	5481173	1	0.50	Antibacterial
Bromfenac	60726	1	0.55	Anti-inflammatory

TABLE 5.9: **Fatty liver drugs predicted using molecular docking studies into the carnitine binding site of the CrAT protein**.

5.10 Discussion

In this study, a novel approach to detect fatty liver drugs using metabolic network based target selection and computational molecular modeling method had been reported. The important targets that play important role in the fatty acid metabolism have been identified using the kinetic modeling of the central hepatic metabolism. Out of the eight top ranked target, CPT1 was selected for further investigation and computational modeling studies. The inhibitors of the CPT1 were selected from the chembl database (ref) and two different hypothesis were proposed. First hypothesis was proposed to predict drugs that could bind to the malonly-CoA binding site of the CPT1 and there by result in no beta-oxidation of the fatty acids. The inhibitors of CPT1 liver isoform in humans were used to create common feature based pharmacophore model. The inhibitors were obtained from the same source (tested in the same assays) and therefore serves as gold standard to generate the model. All the inhibitors had strong activity value as shown in the table 5.2, a close inspection of the structures of these actives inhibitors in the training set reveals all these compounds three common features such as one hydrophobic group and two hydrogen bond acceptor lipid groups with an interfeature distance less than 2.5 Å. On comparison of the functional groups of these actives sets confirms the carboxyl group which denotes the hydrogen bond acceptor lipid are important for the interaction with the enzyme pocket. However, having heterocyclic ring in the structure makes them hydrophobic and which is essential for it to be transported to the site of action by diffusion across membranes. Additionally, it was also noticed that some of inhibitors of the CPT1 have structural similarity with carnitine, supporting the second hypothesis in this study can be true. The 12 known fatty liver drugs were mapped to the common feature pharmacophore model, out of which 5 drugs were mapped as shown in the figure 5.8 . Amiodarone which is one of the classical fatty liver is also a inhibitor of CPT1, this supports our hypothesis that drugs inhibiting CPT1 can results in fatty liver syndrome. Based on the pharmacophore model 100 approved drugs were predicted to have similar pharmacophoric features like that of CPT1 inhibitors with individual fit value. Top 30 drugs were selected having highest fit value. It is commonly observed that binding affinity of drugs to the intended

molecular target is much higher than to their off-targets. Therefore, it may be possible that drugs that are administrated at higher doses at regular interval can induce fatty liver much faster that the drugs that are taken for a short duration and in less quantity. So considering the daily administrative dose as well as therapeutic area, 23 final drugs were selected which are suggested for experimental validation as shown in the table tab:pharmdrugs.

Molecular docking studies of the drugs predicted using the fragments and similarity search method reveals that drugs can also bind to the carnitine-binding site of the protein resulting in competitive inhibition. According to the second hypothesis, three drugs were predicted. Since, the crystal structure of the CPT1 was not available, the structure of human CrAT protein was used for molecular docking studies as shown in the table 5.9. One of the main shortcoming of this investigation is unavailability of the CPT1 structure, it is therefore difficult to predict how exactly the carnitine similar analogues can bind to CPT1 in a real biological system.

5.11 Conclusion

As illustrated above, the approach presented in this chapter is a novel way to determine target, drugs and the mechanism involved in drug-induced fatty liver syndrome. The uniqueness of this approach includes the consideration of different layers of information: selection of targets using a kinetic model of the central hepatic metabolism, generating of models based on the pharmacophores as well as ligand-protein interaction. Beyond the prediction of potential target for fatty liver and the mechanism involved, the modeling approach presented in this study is expected to open up many additional exciting possibilities in the near future, for e.g data on drugs-target interactions may be integrated with the kinetic model to further enhance the model's accuracy.

The combination of computational system biology, structural bioinformatics and cheminformatics based approach as presented in this study could become indispensable tool during the pre-clinical and clinical phases of new-drug development for studying the nature of adverse side-effects.

Chapter 6

Prediction of drugs interacting with HLA alleles

6.1 Introduction

Adverse drugs reactions (ADRs) remain a significant source of patient morbidity throughout the world[104] . ADRs can be broadly classified into two types: such as type A reactions which are often considered as predictable, common and related to the drug's pharmacological activity and type B reactions also known as 'Idiosyncratic drug reactions' which are mostly unpredictable and not related to the drug's pharmacological actions. Type B reactions includes drug intolerance reactions (e.g asthma) and immune mediated adverse drugs reactions (IM-ADRs) also known as drug hypersensitivity reactions[105]. IM-ADRs are among the most difficult type of ADRs to predict. The human leukocyte antigen (HLA) is well known for its association with certain diseases like ankylosing spondylitis, celiac disease and many others. Recently, based on clinical studies it was established such reactions based on immune mechanisms have genetic associations and strong linkages between drug hypersensitivity reactions to several drugs and HLA alleles have been identified[106]. The discovery of new associations between drug toxicities and specific HLA alleles has been well documented for immunologically mediated reactions associated with the use of abacavir in patients expressing the HLA molecular variant B*57:01[107]. HLA- B*57:01, like other HLA alleles, is not equally represented among all ethnic populations as shown in the table 6.1 .

TABLE 6.1: **Allele frequencies in various populations (Meyer et, al., 2007).**

Population	B*57:01 (%)	B*15:02 (%)	B*58:01 (%)
Cuacasian	3.3	<1	0.8
Sub-Saharan African	1	<1	5.8
Han Chinese	<1	10.2	7
Puyuma	<1	18	<1
Okinawan	<1	<1	<1
Singapore	<1	11.6	5.8
Korean	<1	0.5	5.5

The significance of these discoveries has led the European Medicine Agency (EMA), the USA Food and Drug Administration (FDA) and other regulatory agencies to recommend HLA based gene testing before initiation of drug treatment[108] . Abacavir (cyclopropylaminopurinylcyclopentane: ABC) is a structural analogue of guanosine and reverse transcriptase inhibitor which is used in the treatment of human immunodeficiency virus (HIV) infection and the acquired immunodeficiency syndrome (AIDS). It has been reported that in approximately 8 % of the patients, abacavir is associated with significant immune-mediated drug hypersensitivity, which is strongly associated with the presence of the HLA-B*57:01 allele[109] . Abacavir is a deoxy-guanosine base and is metabolized into carbovir triphosphate[110] . Though abacavir is metabolized by the liver, it does not inhibit or induce cytochrome P-450 enzymes and does not interact with medications metabolized by cytochromes[108] .

Three most important complementary models for the immune-mediated severe drugs reactions mechanism have been reported[105]. The pro-hapten (or hapten model) states that chemically inert drug may result in reactive metabolites and therefore induce an immune response by modifying host's self-proteins and forming de novo antigens complexes[111] . The p-i concept which is based on the pharmacologic interactions with immune receptors states that a chemically inert drug can activate T-cell response if they exhibit high affinity to T-cells receptors (TCR) or major histocompatibility complex (MHC) molecules[112]. The danger model states the danger signals such as chemical, physical or viral stress other than drugs are involved in overcoming the immune tolerance barriers[113]. However, the most convincing mechanisms was described by McCluskey group[106] . Human leukocyte antigen class I-restricted activation of CD8 + T cells provides the immunogenetic basis of a systemic drug hypersensitivity.Immunity 28:822–832), which states that the activation of $CD8^+$ cells in a strictly HLA-B*57:01-restricted manner is associated with abacavir induced adverse drugs reactions. Following the mechanism proposed by Illing, et.al[106]. an alternative hypothesis ''altered self-repertoire' hypothesis as a mechanism for drug hypersensitivity is explained by

Peters group [107] . According to Ostrov,et.al [107], the binding groove of HLA-B*57:01 can accommodate abacavir and hence alter the repertoire of self-peptide proteins that are bound and presented to T-cell receptors eliciting immune response.

In the earlier study reported by Ostrov,et.al [107], it was illustrated that abacavir interacts with the residues in the F-pocket of HLA-B*57:01 and is in contact with the side-chain of the C-terminal residue of the bound peptide. This type of binding of abacavir with HLA specific antigen (HLA-B*57:01) results in an alteration of its specificity for self-peptides. In this study, the mechanism proposed by Peters group was considered as the standard mechanism to develop a *in silico* prediction protocol for drugs which are structurally similar to abacavir and may bind to HLA-B*57:01 in a similar manner and therefore could result in immune response.

6.2 Methods

In this study, a protocol has been designed which is based on the Abacavir-HLA-B*57:01 interactions (ligand-receptor). The initial step involves analysis of the key residues involved in the interaction of the abacavir-HLA complex and therefore identifying the pharmacophoric groups. Two-dimensional (2D) and three-dimensional (3D) similarity search were implemented to filter drugs having structural similarities with abacavir. The screened drugs were docked into the binding cavity of HLA-B*57:01 considering abacavir as the standard ligand. The molecular docking study results were analyzed and suggested for experimental validation. The workflow is shown in the figure 6.1.

6.2.1 Analysis of pharmacophoric groups

To identify the key residues involved in the interaction of Abacavir with the HLA complex a pharmacophore based feature mapping was developed using the automatic pharmacophore generation option of the Discovery Studio software. The crystal structure of the HLA-abacavir complex (PDB id: 3UPR) was used and 10 features pharmacophore of ligand (abacavir) was generated. The features includes two hydrogen bond acceptor, six hydrogen bond donor, one hydrophobic and one ring-aromatic. In abacavir, cyclo propyl moiety has hydrogen donor features, cyclopentenyl moiety has both hydrogen donor and hydrophobic features whereas the 2-amino purine moiety contains hydrogen acceptor, hydrogen donor and ring aromatic features as shown in the figure 6.2 .

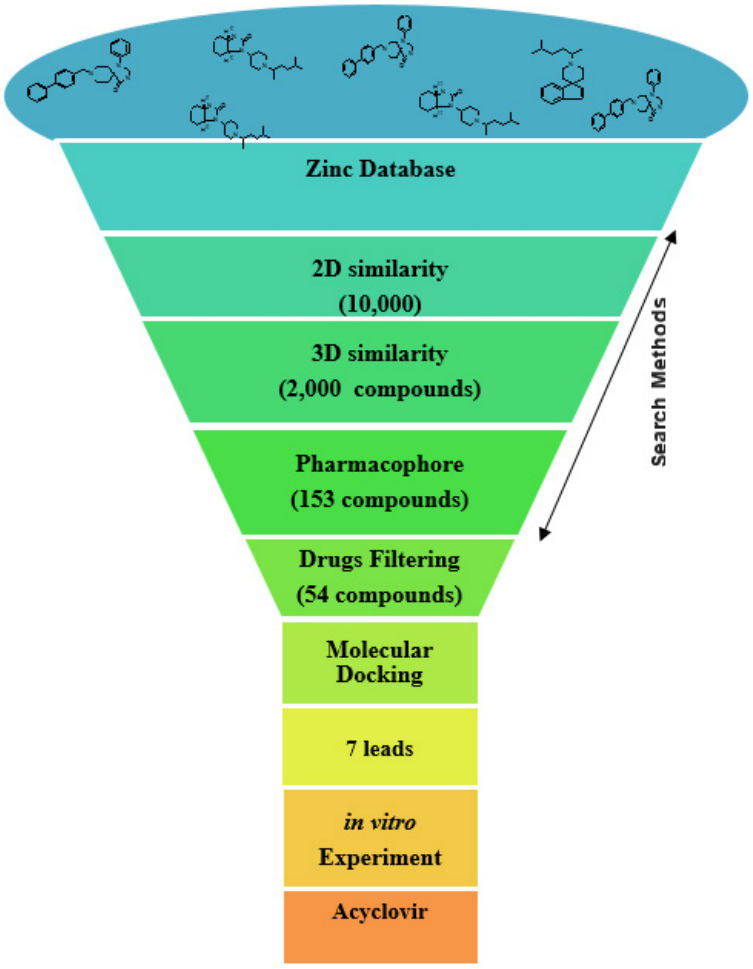


FIGURE 6.1: **A diagrammatic representation of the protocol for prediction [115].**

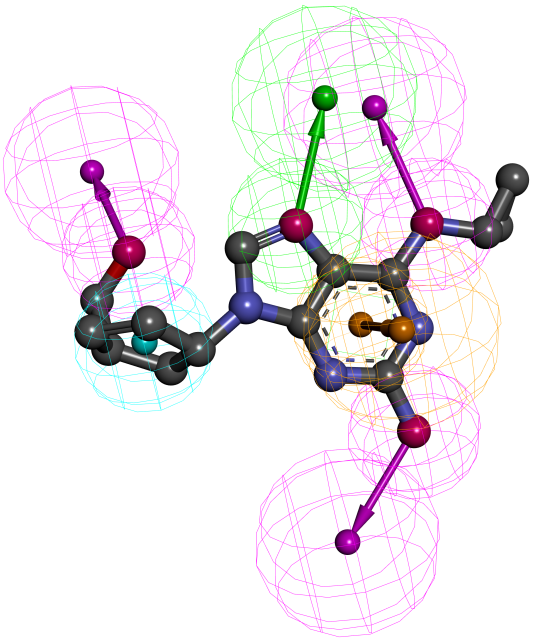


FIGURE 6.2: **Pharmacophoric features of abacavir.** The hydrogen acceptors shown in green, hydrogen donor in magenta, hydrophobic feature in cyan and orange represents the ring-aromatic feature.

6.2.2 Similarity based screening

The structure of the lead molecule abacavir was compared with database molecules based on 2D similarity and 3D similarity. In this study, zinc database was screened using 2D similarity search based on a concatenated fingerprint consisting of the 'FP2' and 'FP4' fingerprints of MyChem as explained in the sections 3.3.1 and 3.4.1 . The structural similarity score was calculated using Tanimoto index[44] , where 0 means no similarity and 1 means maximum similarity. A total of 10,000 molecules were screened considering a threshold of 0.60 and above.

Furthermore, these 10,000 molecules were further compared with abacavir based on their 3D similarity. The 3D similarity was performed using a superimposition algorithm developed previously in the group[114]. The 3D similarity scoring is based on the root mean square deviation (rmsd) function which takes into account the distance between the atoms. This score reflects the degree of superimposition between the query molecule (abacavir) and the database entries. Using the 3D scores with a threshold of rmsd less than 1.5 Å, a total of 2000, compounds were screened.

The purine core of abacavir was observed to be important from the crystal structure of the complex (PDB code: 3UPR). Accordingly, the purine core was marked as one of the chemical features to be retained in the screened drugs. Similarly, the hydroxyl group of abacavir was hypothesized to be actively involved in the binding preference. Moreover, it was also analyzed from the crystal structure that the spatial arrangement of the aromatic ring is an important recognition element. The pharmacophoric features and important functional groups of abacavir involved in the interactions with the receptor were considered to further optimize the screening pipeline which consists of 153 compounds. These 153 compounds were further reduced to 54 compounds by selecting only approved drugs and special preference to antiviral drugs (i,e drugs from the same therapeutic class as abacavir). Finally, with reference to the chemical tractability and the quality of superimpositions, 54 compounds (structurally similar to abacavir) were docked into the binding cavity of the HLA complex.

In the table 6.2 the 2D similarity and 3D similarity scores are reported with the respective RMSD and alignment scores of individual drugs compared to abacavir. These score were recalculated using KNIME while preparing this thesis.

TABLE 6.2: **2D similarity and 3D similarity scores**

Pubchem ID	Name	2D scores	3D scores	RMSD	Alligment scores
441300	Abacavir	1	1	0	126.68
2422	Bohemine	0.77	0.67	0.33	106.70
160355	Roscovitine	0.72	0.65	0.53	96.96
20279	Cladribine	0.64	0.43	0.48	114.40
2022	Acyclovir	0.61	0.36	0.591	95.208
3011155	Arranon	0.61	0.45	0.55	104.30
14978	Sangivamycin	0.60	0.44	0.61	106.09
4201	Minoxidil	0.50	0.40	0.56	88. 913

6.2.3 Molecular docking

Molecular docking was performed to probe the binding sites of the 54 selected drugs into HLA-B*57:01. The 2D structures of the molecules were cleaned and water molecules or salts attached were removed. The 3D coordinates of the molecules were created using the Discovery Studio software . Computational docking was performed using GOLD 5.2 (Genetic Optimization for Ligand Docking) and the GOLDScore scoring function. The crystal structure of HLA-B*57:01 (abacavir-induced MHC complex) having PDB ID: 3UPR was selected as the standard structure for docking. The original ligand (abacavir) confirmation and cavity was used to defined the active binding site (radius 10.0 Å), which also includes the the peptide binding area of the HLA-B*57:01 molecule. GOLD docking software permits full ligand flexibility and partial protein flexibility for a maximum of ten user defined residues including both backbone and side chain residues. All other additional standard parameters were kept default.

6.2.3.1 Analysis of docking results

The docking results were visually inspected using PyMOL(The PyMOL Molecular Graphics System, Version 1.5.0.4 Schröndinger , LLC). The drugs that failed to bind into the defined binding site were ignored and were not considered for further screening. The 2D ligand-interaction diagram for each pose of the drugs were created using Discovery Studio software and interactions with key residues within the binding site of the HLA-B*57:01 molecules were noted. Based on standard ligand interactions and reasonable binding conformations, seven candidates were forwarded as leads for experimental validation.

The drugs roscovitine, cladribine, acyclovir, arranon, minoxodil, sangivamycin and bohemine were selected as final drugs for *in vitro* experiment. All the drugs contained purine core as well as the hydroxyl group important for binding, except minoxidil which was selected as negative control as shown in the figure 6.3 .

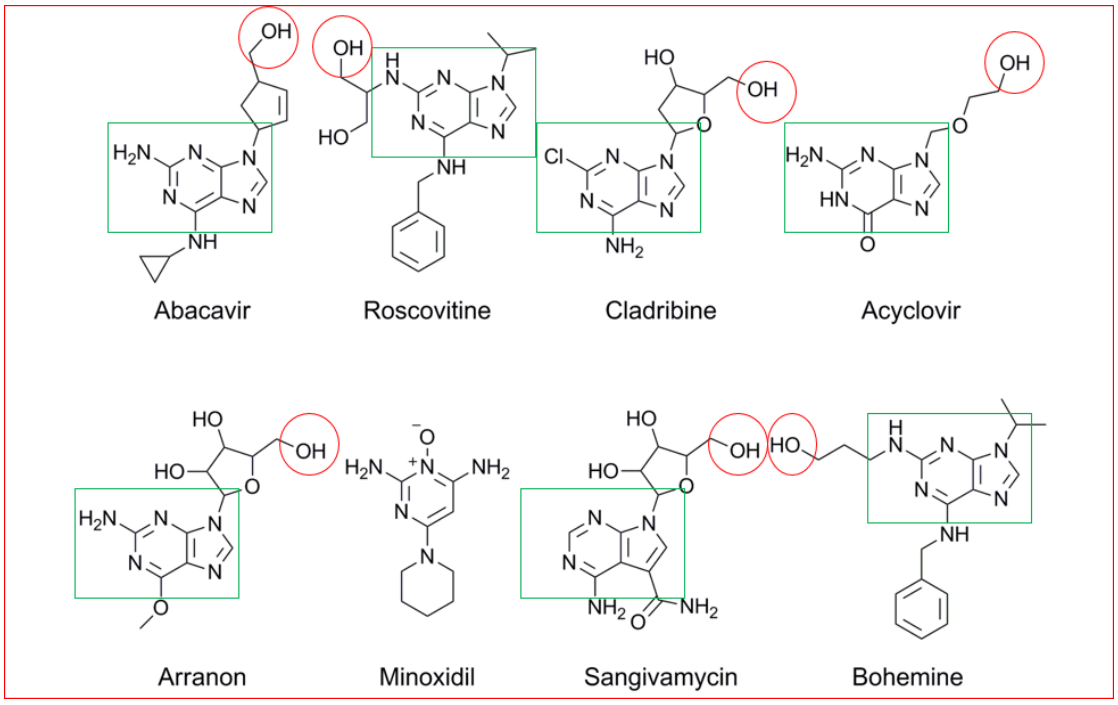


FIGURE 6.3: **Seven drugs as leads obtained after molecular docking studies[115] .** The purine core in highlighted in green box and the hydroxyl group in red circle. The picture was taken from the publication

6.3 Results

Computational identifications of approved drugs which are structurally similar to abacavir was based on the assumption 'structurally similar drugs should have similar biological activities'. Additionally, further layers of information like pharmacophoric groups and binding interactions with the HLA-B*57:01 molecule were also added into the prediction pipeline. The prediction pipeline includes screening of ZINC database based on 2D and 3D similarity-based method followed by pharmacophore based filtering. In order to reduce the huge chemical space of ZINC database, a threshold of 0.60 Tanimoto score was considered. Based on the chemical tractability and reasonable binding interactions, the enriched subset was docked into the HLA- B*57:01 active site using computational docking. Docking results were visually inspected and analyzed in detailed to screen the final seven drugs.

Previous studies[107] revealed that abacavir binds to peptide binding groove of HLA- B*57:01 interacting with the Ser166, Asp 114, Ile 124 and Tyr74 of the HLA molecule. Additionally,

it was observed that the purine core of abacavir plays a crucial role in the molecular recognition. Therefore, all the information was considered for pose selection from the molecular docking studies.

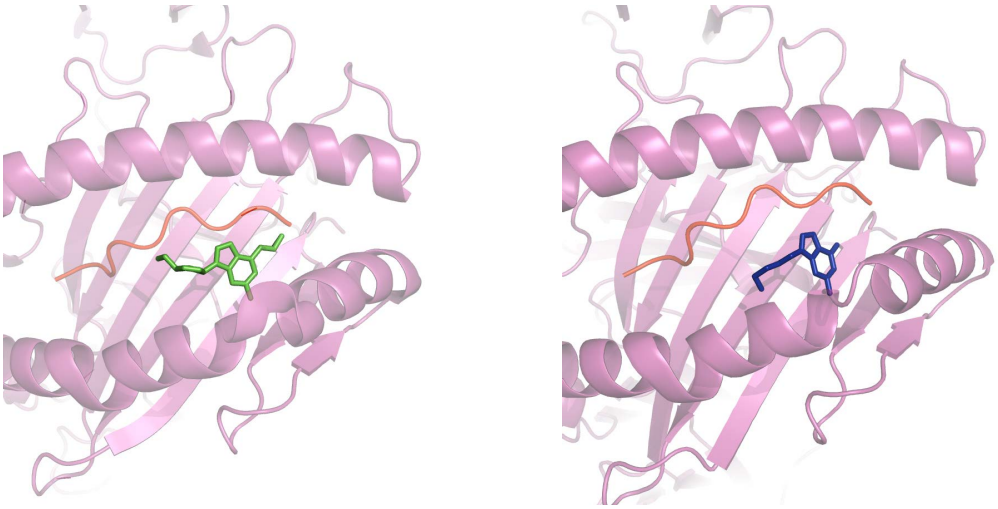
Visual inspection of the docking poses in terms of the ligand-receptor interactions, maximum similarity between the reference structure (abacavir) and the database entries as well as the occupancy and conformation of the screened drugs into the defined active site of the HLA- B*57:01 molecule led to the prediction of seven drugs as candidates for experimental studies.

Drugs	**Therapeutic class**
Roscovitine	antineoplastic
Cladribine	antineoplastic
Acyclovir	antiviral
Arranon	antineoplastic
Minoxodil	vasodilator
Sangivamycin	antibiotic
Bohemine	kinase inhibitor

6.3.1 Computational validation

The computational docking studies revealed that acyclovir binds to HLA- B*57:01 molecule similarly to abacavir. Like abacavir, acyclovir too occupies the floor of the peptide binding groove of HLA- B*57:01 molecule as shown in the figure 6.4 . This kind of arrangement is important in order to increase the affinity of certain self peptides resulting in an altered peptide-binding repertoire as identified in the previous study[115].

Furthermore, analysis of the ligand-receptor interaction diagram of acyclovir- HLA- B*57:01 complex obtained from docking studies shows that acyclovir interacts with the same residues of the HLA- B*57:01 molecule as observed in abacavir- HLA- B*57:01 complex. Acyclovir forms hydrogen bond interactions with the side-chains of Ser 116, Asp 114 and Tyr118 of HLA- B*57:01 molecule. Additionally, there is a Pi interaction with Trp 147, Tyr74 of HLA-B*57:01 molecule as shown in the figure 6.5 .



(A) Abacavir binding to HLA- B*57:01 (B) Acyclovir binding to HLA- B*57:01

FIGURE 6.4: **Computational identification of binding site for acyclovir in HLA- B*57:01 similar to abacavir** Abacavir (left) shown in green and acyclovir (right) shown in blue binding to the F-pocket of HLA- B*57:01 which are typically in contact with the side chain of the C-terminal residue of the bound peptide shown in orange [115] .

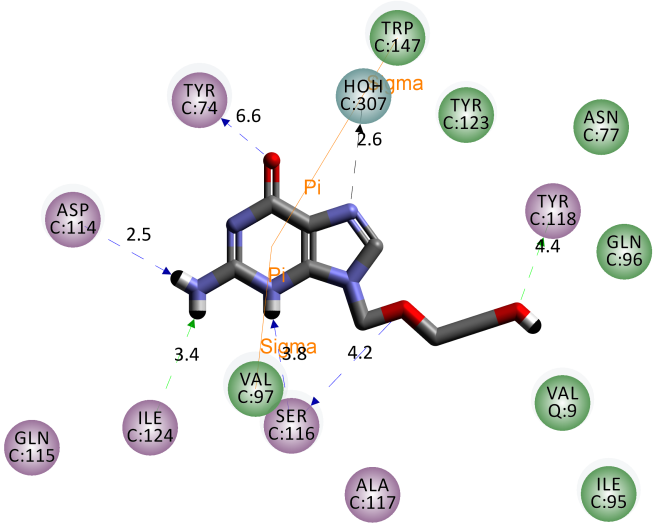
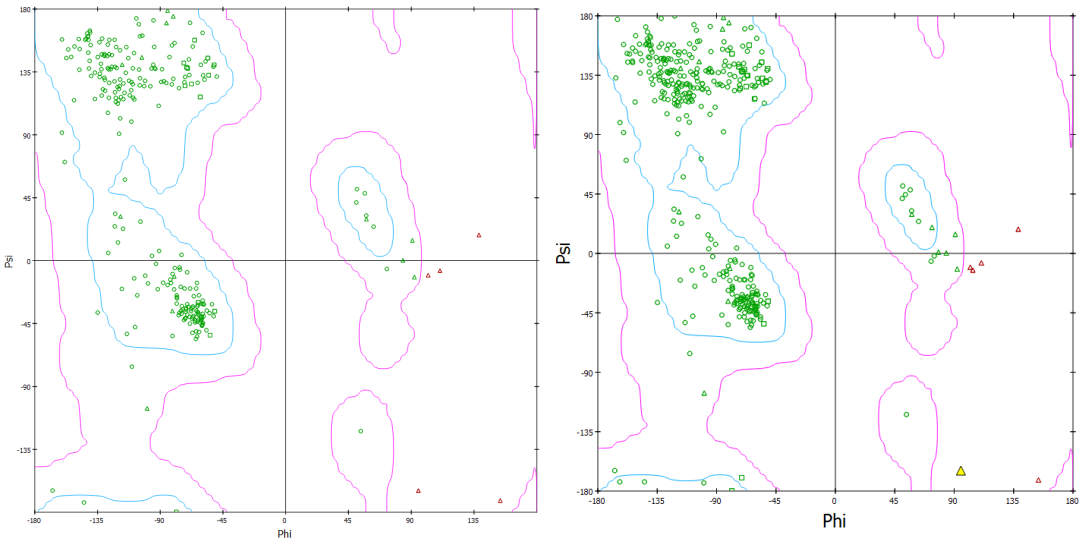


FIGURE 6.5: **Ligand interaction diagram of acyclovir-HLA- B*57:01 complex.** Hydrogen bond interactions between the Ile 124 (MHC molecule) main chains and acyclovir is represented by green dashed arrow. Hydrogen bond interactions between side-chains of Ser 116, Asp 114 (MHC molecule) and acyclovir is represented by blue dashed arrow. Pi interactions between the Trp147, Tyr74 of the MHC molecule and acyclovir is shown in orange line.

In addition, examination of the Ramchandran plot [116] of the back bone angles of the crystal structure of HLA- B*57:01 molecule before and after molecular docking with acyclovir reveals that they both fall into the commonly observed regions of the psi-phi space as shown in the figure 6.6 . The percentage of residues in the favored region and allowed region is 96.9 % and 3.1 % respectively, for both the structures. Whereas, there were no residues in the outlier

region. This indicates the absence of any steric clashes or distortion in the protein structure after docking of acyclovir into the peptide binding groove of HLA- B*57:01 molecule.



(A) Protein report of the crystal structure

(B) Protein report after molecular docking

FIGURE 6.6: **Ramchandran plot for HLA- B*57:01 molecule before and after docking with acyclovir**. The most favored and favoured region are indicated with blue and pink colors, respectively. The disallowed region is in white color.

6.3.2 Experimental validation

In the experimental analysis, each of the seven predicted drugs were tested in order to confirm if their presence could enhance the HLA- B*57:01 binding affinity for three 9-mer peptides with a C-terminal valine, which is known to bind HLA- B*57:01 with high affinity only in presence of abacavir. This analysis revealed that acyclovir has consistent increase in HLA-B*57:01 affinity in presence of these three peptides, while the other six drugs have smaller or no effects at all. Though it was found that acyclovir interacts with the HLA- B*57:01 molecule in a similar manner as abacavir, it was suggested that under *in-vitro* conditions, acyclovir does not have the capacity to stimulate T-cells. The experimental studies were carried out by Peters and group, at Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, California, United States of America. Further information on the experimental studies and procedure can be obtained from the published work[115] .

6.4 Conclusion

Identification of drugs candidates associated with IM-ADRs is an important step in drug design and development. Peters and group[107] recently proposed a novel mechanism through

which abacavir interacts with HLA- B*57:01 molecule leading to an alteration of the specificity of HLA- B*57:01 for self-peptides and associated immune mediated T-cell hypersensitivity reaction . This hypothesis is termed as 'altered self-repertoire' hypothesis. In this study, this hypothesis was used to screen drugs that could have potential effects on the self-peptide repertoire. *In silico* prediction method was developed considering abacavir as the lead structure and to predict drugs that are structurally similar to abacavir. Since, the crystal structure of abacavir-HLA- B*57:01 complex was available, screened drugs were further investigated based on molecular docking experiments. Six drugs which were structurally similar to abacavir and one drug (minoxidil) having similar pharmacophric groups but structurally less similar to abacvir; were suggested for *in vitro* experiments. Of the panel of seven drugs tested in the *in vitro* studies, acyclovir was found to have the strongest effect and alter the binding specificity and ligand repertoire of HLA- B*57:01 molecule. This effect produced by acyclovir was in a qualitatively similar manner as abacavir, however quantitatively to a much lower degree. Acylcovir is structurally similar to abacavir, also a guanosine analogue, and shares a Tanimoto similarity score with abacavir above 0.80. They both are from the same therapeutic area (i.e antiviral agent).

In this study, it was possible to established an *in silico* screening protocol taking into account the structural similarity measure, pharmocophoric features and supported by computational validation which lead to a novel finding that drug acyclovir that binds to peptide binding groove of HLA- B*57:01 molecule.

In the *in vitro* experiments, quantitatively the presence of acyclovir led to 2-3 fold increases in binding affinity as measured in IC_{50} where as abacavir led to a 1,000 fold increase in IC50 values for individual peptides and large number of peptides presented exclusively in the presence of the drug. Acyclovir have a safety profile and the effects of the magnitude observed here are below the threshold to induced IM-ADRs. However, this study establishes a first estimate of a threshold of what can be considered a safe in the experimental essays [115].

Chapter 7

Summary

7.1 Summary

Exposure to various chemicals agents through cosmetics, medications, preserved food, environments and many other sources have resulted in serious health issues in humans. Additionally, regulatory authorities from Europe and United States of America have recognized the risk associated with combined exposure to multiple chemicals. Testing all possible combinations of these thousands of compounds is impractical and time consuming. The main aim of the thesis is to address the problem of off-targets effects of chemical structures by applying and developing cheminformatics, structural bioinformatics and computational systems biology approaches.

This dissertation is divided into four main projects representing four different computational methods to aid different level of toxicological investigations.

In project I (chapter 3) a novel ensemble approach based on the structural similarity of the chemical compounds and identifications of toxic fragments was implemented to predict rodent oral toxicity. This approach showed powerful and consistently best performance over available commercial as well as other public methods.

In project II (chapter 4) different machine learning models were developed and compared using Tox 21 challenge 2014 data, to predict the outcomes of the compounds that have the potential to interact with the targets active in toxicological pathways. The methods proposed in this study can be used by the regulatory agencies to access the toxicity of these target specific compounds in large scale.

In project III (chapter 5) a novel approach integrating the trio concept of 'computational system biology, cheminformatics and structural bioinformatics' to predict drugs induced

metabolic syndrome have been described. Two different novel mechanisms for drug induced fatty liver syndrome has been proposed and computationally validated.

In project IV (chapter 6) a *in silico* screening protocol was established taking into the structurally similarity, pharmacophoric features and validation using computational docking studies. This approach led to the identification of novel binding site for acyclovir in the peptide binding groove of the human leukocyte antigen (HLA) specific allele. Such specific binding of drugs to the specific HLA alleles results in immune-mediated adverse drug reactions.

The over all contribution of the thesis includes development of different computational methods that can successfully be applied to address the problems of off-targets effects at various levels of system inference.

7.2 Discussion and conclusion

This thesis is focused on the development of *in silico* methods for prediction of toxic outcomes. The thesis envisages a global view in response to toxicity profiles of chemical compounds as well as local (specific) consideration on the mechanism and pathway level. For example, a drug or a chemical compound might interact with a molecular target which can result in interactions with multiple molecular targets including both therapeutic as well as off-target with different affinities. In this process, consequently it can activate different signaling pathways or interact with functional pathways. Additionally, such interactions at cellular level can produce toxic effects on certain organs. This can be further extended to the ADRs profile of population sharing similar toxicological pathways or network.

The study reported in the chapter 3 establishes a novel method for prediction rodent toxicity only available information of chemical data. Similarity method needs little information to formulate a reasonable query; specifically nothing need to be known about the active confirmation of the query molecule as its based on the 2D coordinates of the molecules. Similarity-based model can be really useful when there is little or no information on the target is known. Implementation of similarity method are computationally inexpensive and large data sets can be predicted and analyzed in less time. A major limitation of this approach is that prediction for each query (unknown) molecules depends largely on the activity and diversity of the structurally similar molecules present in its training set which accounts on the neighborhood behavior of the molecules. Though the molecules are compared using structural molecular fingerprints, however the scoring is based on the global similarity between the compared molecules. Most of times it is observed that the property of a chemical compounds is deeply connected with a local feature present in it. Hence, understanding this local patterns is extremely important in order to understand the activity of a molecule.

In this study, fragmentation approach was implemented inspired by the 'local feature based association' hypothesis. It was observed in this study, that addition of fragments based identification improved the prediction for the most toxic class I by 5 % , class II by 9 % and class III by 3 % on cross-validation set as shown in table 3.3 . Based on only individual fingerprints the prediction method achieved highest performance with ECFP4 , followed by a combination of FP24 and fragments and only FP24 fingerprints individually has the least performance on cross-validation set. However, the best performance on external validation was achieved by combination of FP24 and fragments. The ensemble method developed in this study was compared with the commercial software TOPKAT and publicly available QSAR based method T.E.S.T and have outperformed in all evaluation measures on the external set as shown in table 3.3 .

Chapter 4 represents an ensemble model based on similarity and fragments approach was developed to predict the toxicity of the chemicals by using concatenated 'FP24' fingerprints and statistically analyzed toxic fragments. The similarity approach is based on the neighborhood behavior of chemicals compound and representation of their structural feature using molecular fingerprints. In the fragment based approach, each molecular structure was compared to statistically analyzed fragments by using substructure search that represents the presence or absence of particular toxic substructures in the molecules with a confidence score. The application of this ensemble approach has the potential to achieve a classification model with high prediction accuracy scores as well as prediction confidence. The major advantage of this approach is the capability to incorporate addition of new toxicity data easily and developing models for other toxic end points. In this study, two different methods were proposed for the prediction of interference of the Tox21 challenge data set consisting of chemical compounds in two major biological pathways; nuclear receptor pathway and stress response pathway. The data was generated in a standard uniform experimental setup which serves as a gold standard data source for evaluating performance of different prediction methods. In the first section, a similarity-based fingerprint method is reported which is based on the *' similarity property principle'* and concatenated fingerprints with molecular descriptors based binary fingerprints. This method performed relatively better in combination of different fingerprints and descriptors. In the second section, machine learning based models were developed in combination of individual fingerprints and as well as in combination with molecular property based descriptors. It is observed that the machine learning models based on RF classifier achieved best performance when compared to other two machine learning models (NB and PNN) as well as the similar-based fingerprint method. The superior performance of the RF models can be attributed to the different tunning parameters chosen for individual targets (i,e RF algorithm is robust to different tuning parameters). On the other hand, the poor performance of the PNN model can be explained by its strong inclination towards the majority class coverage of the training set. To understand this behavior

in detail, the results were analyzed and it is observed that PNN models were able to correctly predict all the true negatives in the external validation with a confidence score higher than 0.9 but failed to correctly predict the actives (true positives) for all the three targets. On the other hand, NB models could predict the highest number of true positives with confidence scores higher than 0.99 in comparison to the RF and PNN based models. However, NB lacks the prediction ability when it comes to true negatives (inactives /majority class). Taking an example of a result from randomly selected target ER-lLBD, this trend was analyzed in details for each models taking in consideration of two different fingerprints (MACCS; ECFP4) as shown in table 4.16 . RF based models are able to identify the patterns associated with respective classes in case of imbalanced data set.

Furthermore, it is toxicity alert encoded fingerprints (ToxPrint) as well as atom state based fingerprints (EState) failed to obtain consistent performance across most of the targets for various models. This could be due to the fact the chemotypes in the training set do not match with the pre-defined toxicity alerts encoded in the ToxPrint fingerprints. On the other hand the MACCS fingerprints encodes the similar substructures like the one present in the chemical space of the Tox21 challenge data and therefore performed better and consistent across various machine learning models. This supports the fact that prediction of toxicity can not always be encountered using a global approach (i,e identification of presence of certain toxic alerts in the chemical space). Target specificity and local substructures closely associated with the chemical space addressed in the study plays an important role in the prediction process. In addition, selection of optimal descriptors which could represent the chemical space and an unbiased classifier which can learn the patterns and used this knowledge to predict activity of an unknown compound is in real a true essence of predictive science.

Overall, in this study it can be emphasized that a simple RF based classifier consistently demonstrated robust prediction for all the three targets. This method is relatively simpler and computationally less expensive than other methods of the Tox21 challenge. The results of this study are equally good when compared with the Tox21 challenge top models with respect to AUC values and are better with respect to balance accuracies. This further adds to the usability of the optimal method developed in this study. The methods will be publicly available. In this study, a random forest based *in silico* toxicity prediction method is reported, emphasizing the importance of predictive toxicology as a fast and reliable way to predict the toxic outcome of chemical compounds. Additionally, it can be concluded from the result of this study that the combination and application of RF algorithm and substructure based molecular fingerprints (MACCS) can be regarded as a very promising prediction tool for evaluation of toxic effects of new chemicals.

In chapter5, a novel approach to detect fatty liver drugs using metabolic network based target selection and computational molecular modeling method had been reported. The important

targets that play important role in the fatty acid metabolism have been identified using the kinetic modeling of the central hepatic metabolism. Out of the eight top ranked target, CPT1 was selected for further investigation and computational modeling studies. The inhibitors of the CPT1 were selected from the chembl database (ref) and two different hypotheses were proposed. First hypothesis was proposed to predict drugs that could bind to the malonly-CoA binding site of the CPT1 and there by result in no beta-oxidation of the fatty acids. The inhibitors of CPT1 liver isoform in humans were used to create common feature based pharmacophore model. The inhibitors were obtained from the same source (tested in the same assays) and therefore serves as gold standard to generate the model. All the inhibitors had strong activity value as shown in the table 5.2, a close inspection of the structures of these actives inhibitors in the training set reveals all these compounds three common features such as one hydrophobic group and two hydrogen bond acceptor lipid groups with an interfeature distance less than 2.5 Å. On comparison of the functional groups of these actives sets confirms the carboxyl group which denotes the hydrogen bond acceptor lipid are important for the interaction with the enzyme pocket. However, having heterocyclic ring in the structure makes them hydrophobic and which is essential for it to be transported to the site of action by diffusion across membranes. Additionally, it was also noticed that some of inhibitors of the CPT1 have structural similarity with carnitine, supporting the second hypothesis in this study is valid. The 12 known fatty liver drugs were mapped to the common feature pharmacophore model, out of which 5 drugs were mapped as shown in the figure 5.8 . Amiodarone which is one of the classical fatty liver is also a inhibitor of CPT1, this supports the hypothesis that drugs inhibiting CPT1 can results in fatty liver syndrome. Based on the pharmacophore model 100 approved drugs were predicted to have similar pharmacophoric features like that of CPT1 inhibitors with individual fit value. Top 30 drugs were selected having highest fit value. It is commonly observed that binding affinity of drugs to the intended molecular target is much higher than to their off-targets. Therefore, it may be possible that drugs that are administrated at higher doses at regular interval can induce fatty liver much faster that the drugs that are taken for a short duration and in less quantity. So considering the daily administrative dose as well as therapeutic area, 23 final drugs were selected which are suggested for experimental validation as shown in the table tab:pharmdrugs.

Molecular docking studies of the drugs predicted using the fragments and similarity search method reveals that drugs can also bind to the carnitine-binding site of the protein resulting in competitive inhibition. According to the second hypothesis, three drugs were predicted. Since, the crystal structure of the CPT1 was not available, the structure of human CrAT protein was used for molecular docking studies as shown in the table 5.9. One of the main shortcoming of this investigation is unavailability of the CPT1 structure, it is therefore difficult to predict how exactly the carnitine similar analogues can bind to CPT1 in a real biological system.

As illustrated above, the approach presented in this chapter is a novel way to determine target, drugs and the mechanism involved in drug-induced fatty liver syndrome. The uniqueness of this approach includes the consideration of different layers of information: selection of targets using a kinetic model of the central hepatic metabolism, generating of models based on the pharmacophores as well as ligand-protein interaction. Beyond the prediction of potential target for fatty liver and the mechanism involved, the modeling approach presented in this study is expected to open up many additional exciting possibilities in the near future, for e.g data on drugs-target interactions may be integrated with the kinetic model to further enhance the model's accuracy. The combination of computational system biology, structural bioinformatics and cheminformatics based approach as presented in this study could become indispensable tool during the pre-clinical and clinical phases of new-drug development for studying the nature of adverse side-effects.

In the study reported in chapter 6, it was possible to established an *in silico* screening protocol taking into the structural similarity measure, pharmacophoric features and supported by computational validation which lead to the identification of novel binding of antiviral drug acyclovir to peptide binding groove of HLA- B*57:01 molecule.

Identification of drugs candidates associated with IM-ADRs is an important step in drug design and development. Peters and group [107] recently proposed a novel mechanism through which abacavir interacts with HLA- B*57:01 molecule leading to an alteration of the specificity of HLA- B*57:01 for self-peptides and associated immune mediated T-cell hypersensitivity reaction . This hypothesis is termed as 'altered self-repertoire' hypothesis. In this study, this hypothesis was used to screen drugs that could have potential effects on the self-peptide repertoire. *In silico* prediction method was developed considering abacavir as the lead structure and to predict drugs that are structurally similar to abacavir. Since, the crystal structure of abacavir-HLA- B*57:01 complex was available, screened drugs were further investigated based on molecular docking experiments. Six drugs which were structurally similar to abacavir and one drug (minoxidil) having similar pharmacophoric groups but structurally less similar to abacavir; were suggested for *in vitro* experiments. Of the panel of seven drugs tested in the *in vitro* studies, acyclovir was found to have the strongest effect and alter the binding specificity and ligand repertoire of HLA- B*57:01 molecule. This effect produced by acyclovir was in a qualitatively similar manner as abacavir, however quantitatively to a much lower degree. Acyclovir is structurally similar to abacavir is also a guanosine analogue and share a Tanimoto similarity score with abacavir above 0.80. They both are from the same therapeutic area (i.e antiviral agent). In the *in vitro* experiments, quantitatively the presence of acyclovir led to 2-3 fold increases in binding affinity as measured in IC50 where as abacavir led to a 1,000 fold increase in IC50 values for individual peptides and large number of peptides presented exclusively in the presence of the drug. Acyclovir have a safety profile and the

effects of the magnitude observed here are below the threshold to induced IM-ADRs. However, this study establishes a first estimate of a threshold of what can be considered a safe in the experimental essays.

7.3 Limitations of the methods presented in this thesis

Structural-cheminformatics based studies are highly reliant on the existing data. The methods reported in chapter 3, is based on the chemical structures similarity. This makes it completely depended on the chemical space used in the study. The performance of the individual toxicity classes are dependent on the available data eg. toxicity classes (III , IV, V) had better prediction when compared to toxicity classes (I, II, VI) which had limited data. Toxicity predictions from chemical structure is only possible, when the structures are directly connected to know mechanism of action. However, if the toxic effect is the result of several different mechanism working sequentially or simultaneously, then the reliable prediction based on the chemical structure could be difficult. The study can only predict the presence of structural alerts in the compound, further investigation regarding the biological processes is required.

In chapter 4, the machine learning methods are highly dependent on the data set. The molecular descriptors selected in the method development was based on defined chemical space for specific targets. Moreover, the imbalanced data set is another limitations in order to increase the predictive performance of the models.

Chapter5 represents a novel approach to understand the mechanism involved in fatty liver disease, however the nonavailability of the crystal structure of the CPT1 makes it impossible to study in details how the suggested drugs will bind to the CPT1 in a biological system. Therefore, lack of data forms a limitation on the applied methods by reducing the applicability domain of the generated models.

In chapter 6, the identification of novel binding of antiviral drug acyclovir to peptide binding groove of HLA- B*57:01 molecule was discovered. This supports that the drugs from same therapeutic classes can have similar biological activity. However, acyclovir could not produce hypersensitivity reaction as it is in case of abacavir. Further studies are required to investigate what kind of functional groups present in abacavir, upon interaction with HLA- B*57:01 molecule; produces T-cell response. This will help to identify more drugs in the same or different therapeutic class, that can bring the similar adverse drug reactions.

7.4 Perspectives and future work

As highlighted in previous chapters, the methods developed in this thesis has proven as valuable supplements to investigate the toxic response of chemical structures. Given the challenges the regulatory toxicology is currently facing (as explained in chapter3 and chapter4) applications of methods developed in this thesis might prove useful when deciding which chemicals to test thoroughly in classical toxicity test and which end points to investigate. Additionally, the hypotheses proposed in chapter 5 to explain the underlying mechanism behind drug induced fatty liver will help to design better drugs that could avoid such off-targets effects; or help to understand how particular drug-target interactions in a metabolic network can result in metabolic syndrome. In chapter 6, a new threshold was defined to predict the drug induce ADRs related to specific HLA alleles, which can be useful in the process of HLA testing.

These methods can provide the predicted toxicity profiles of the unknown chemicals, that can be later validated *in vitro*. Once the actual profile established this information can be integrated into the *in silico* model.

Finally, there is certainly a need of monitoring of predicted off-target effects of drugs or chemical structures,on a more dose dependent manner, such as in what concentration of the drugs such off-targets effects are triggered. Thus, supporting the saying of Paracelsus that *"Poison is in everything, and no thing is without poison. The dosage makes it either a poison or a remedy"*.

List of Figures

1.1 **Different layers of information considered in this thesis to address off-targets effects** The figure is inspired from [6] and represents drugs interacting with different targets, can effects several pathways resulting in off-targets effects on specific organs, sometimes leading to specific ADRs profile. 3

3.1 **Ensemble method** The workflow represents the data source as well as toxicity classes used in this study. The integration of the global similarity scores and local fragment based approach used for the toxicity prediction of the chemical compounds. 19

3.2 **Illustration of a *k*NN classification model** For k=1, the classification of the yellow query instance as a member of the red class; for k=3 it will be again assigned to the red class based on a 2-1 vote. However in cass of k=5, the nearest neighbours are both green, the model will classify it as a part of the green class with a 3-2 vote. 20

3.3 **ProTox web-server** Given the 2D coordinates of a compound, the web-server calculates the toxicity prediction of the compound using the ensemble approach and present the results with the predicted class, average similarity with the toxic compound in the training set as well as percentage of prediction accuracy. Additionally, if any toxic fragments are identified in the input compound, the fragment is reported with a confidence score. 25

3.4 **ProTox web-server** Given the 2D coordinates of a compound, the web-server calculates the toxicity prediction of the compound and highlights the toxic fragments with individual confidence score. 26

3.5 **Super Natural database** Given the 2D coordinates of a compound, the calculates the toxicity prediction of the compound using the ensemble approach and present the results with the predicted class, as well as properties of the molecule and further possibility to search pathways. 31

4.1 **Concatenated fingerprint** A combination of sub structural fingerprint (MACCS), toxic alerts based fingerprint (ToxPrint), circular fingerprint (ECFP4) and property-based binary fingerprint [64] 38

4.2 **Workflow of the similarity-based fingerprint method** Schematic representation of the similarity based method models used to predict the outcomes of Tox21 data 41

4.3 **Workflow of the machine learning methods** Schematic representation of three different algorithms developed for the predictions of active compounds in three different target classes respectively 44

4.4 **Chemical space analysis** The above figure shows the different actives present in the external set of ER-LBD. The compounds highlighted in pink boxes and blue boxes were correctly predicted by Random Forest classifier and Naïve Bayes classifiers respectively. Additionally, respective confidence scores for each classifier are shown. 50

5.1 **A schematic representation of the process of fatty acid oxidation** The above figure shows different enzymes involved in the fatty acid oxidation process and their respective influences in the process. 59
5.2 **Structure of inhibitors of CPT1 similar to carnitine**. 61
5.3 **Structure of ligands in the training set** Most of the active compounds are large molecules and has similar functional groups 63
5.4 **Structures of the test set**. The compounds similar to carnitine and smaller in size are considered as inactive 64
5.5 **Common pharmacophoric features present in the training set** The negative ionisable feauture is shown in blue, hhydrophobic feature is cyan and ring aromativ feature in orange 65
5.6 **Pharmacophore alignment with the active training set molecules** 67
5.7 **Pharmacophore alignment with the least active training set molecules** 67
5.8 **Common pharmacophoric features of Hypo1 to the known fatty liver drugs**, The hydrophobic feature is indicated in cyan and hydrogen bond acceptor lipid shown in green. Fatty liver drugs represented in this set are amiodarone, methrotrexate, tamoxifen and diltiazem 68
(a) 69
(b) 69
(c) 69
(d) 69
5.9 **Carnitine binding interaction with CrAT protein** 72
5.10 **Cefepime binding interaction with CrAT protein** 72
5.11 **Ceftazidime binding interaction with CrAT protein** 73
5.12 **Bromfenac binding interaction with CrAT protein** 73

6.1 **A diagrammatic representation of the protocol for prediction[115].** 80
6.2 **Pharmacophoric features of abacavir.**The hydrogen acceptors shown in green, hydrogen donor in magenta, hydrophobic feature in cyan and orange represents the ring-aromatic feature. 80
6.3 **Seven drugs as leads obtained after molecular docking studies[115] .** The purine core in highlighted in green box and the hydroxyl group in red circle. The picture was taken from the publication 83
6.4 **Computational identification of binding site for acyclovir in HLA- B*57:01 similar to abacavir** Abacavir (left) shown in green and acyclovir (right) shown in blue binding to the F-pocket of HLA- B*57:01 which are typically in contact with the side chain of the C-terminal residue of the bound peptide shown in orange[115] 85

6.5 **Ligand interaction diagram of acyclovir-HLA- B*57:01 complex.** Hydrogen bond interactions between the Ile 124 (MHC molecule) main chains and acyclovir is represented by green dashed arrow. Hydrogen bond interactions between side-chains of Ser 116, Asp 114 (MHC molecule) and acyclovir is represented by blue dashed arrow. Pi interactions between the Trp147, Tyr74 of the MHC molecule and acyclovir is shown in orange line. 85

6.6 **Ramchandran plot for HLA- B*57:01 molecule before and after docking with acyclovir**. The most favored and favoured region are indicated with blue and pink colors, respectively. The disallowed region is in white color. 86

List of Tables

3.1 **Toxicity classes with definition and LD $_{50}$ value range**. 17
3.2 **Number of compounds per toxicity class in the training set** 17
3.3 **Performance of the prediction method in leave-one-out cross validation using different features** 23
3.4 **Performance of the external validation set** 23
3.5 **Comparison of the prediction method with TOPKAT and T.E.S.T on the external validation set** 24

4.1 **Training set class distribution**. The table shows the total number of compounds as well as number of actives and inactives compounds for each target. 37
4.2 **External test set class distribution**. The table shows the total number of compounds as well as number of actives and inactives compounds for each target. 37
4.3 **Types of fingerprints and their encoding description**. Five different fingerprints used in this study and their respective encoding parameters. 39
4.4 **Molecular descriptors used in the method in combination with fingerprints** . 40
4.5 **Cross validation results for Similarity based prediction**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 46
4.6 **External validation results for Similarity based prediction**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 46
4.7 **Cross validation results for Naive Bayes model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 46
4.8 **External validation results for Naive Bayes model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 47
4.9 **Cross validation results for Random Forest model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 47
4.10 **External validation results for Random Forest model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 48
4.11 **Cross validation results for Probablistic Neural Network model**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 48
4.12 **External validation results for Probablistic Neural Network**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis 49
4.13 **ER-LBD active compounds correctly predicted in External set using RF and NB models using MACCS and ECFP4 fingerprints. alone with confidence scores**. Color denotes different molecules illustrated in the figure 4.4 51
4.14 **External validation result comparison with Tox21 Challenge winners**. Area under the curve (AUC) from receiver-operating characteristics (ROC) analysis . 52
4.15 **External validation result comparison with Tox21 Challenge winners**. Balanced accuracies of the different models and targets 52

4.16 **Classifications of actives and inactives in external set by different models for ER-LB.** 53

5.1 **Top 10 ranked enzymes from the kinetic model of central hepatic metabolism** 58
5.2 **Training set for pharmacophore model.** The table shows the total number of training set compounds and their activity value 62
5.3 **Test set for pharmacophore model.** The table shows the total number of test set compounds and their activity value 63
5.4 **The result of top 10 hypotheses generated by HipHop program.** Three features were found to be most common on the active compounds in the training set, one feature corresponds to 'hydrophobic' (Z) and two features corresponding to 'hydrogen bond acceptor lipid' (H). The higher ranking scores indicates the lesser the probability of chance correlation. The best hypothesis is indicated with the highest value 66
5.5 **Validation of the all the ten hypotheses on an external validation set.** 66
5.6 **Screened drugs based on fragment and similarity search scoring.** 69
5.7 **Top 30 screened drugs with highest fit values predicted by model based on Hypo1.** 71
5.8 **Top 30 screened drugs with highest fit values predicted by model based on Hypo1.** 74
5.9 **Fatty liver drugs predicted using molecular docking studies into the carnitine binding site of the CrAT protein.** 75

6.1 **Allele frequencies in various populations (Meyer et, al., 2007).** 78
6.2 **2D similarity and 3D similarity scores** 82

Abbreviations

AUPRC	**A**rea **U**nder the **P**recision **R**ecall **C**urve
AUROC	**A**rea **U**nder the **R**OC **C**urve
FP	**F**alse **P**ositive
ROC	**R**eceiver **O**perator **C**haracteristic
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
NB	**N**äive **B**ayes
RF	**R**andom **F**orest
PNN	**P**robabilistic **N**ueral **N**etwork
NP	**N**atural **P**roduct
HLA	**H**uman **L**uecocyte **A**ntigen
ADRs	**A**dverse **D**rug **R**eactions
2D	**Two D**imensional
3D	**Three D**imensional

Bibliography

[1] D. C. Liebler and F. P. Guengerich. Elucidating mechanisms of drug-induced toxicity. *Nat Rev Drug Discov*, 4(5):410–420, May 2005.

[2] U Gundert-Remy, H Barth, A Bürkle, GH Degen, and R Landsiedel. Toxicology: a discipline in need of academic anchoring–the point of view of the german society of toxicology. *Archives of toxicology*, 89(10):1881—1893, October 2015. ISSN 0340-5761. doi: 10.1007/s00204-015-1577-7. URL http://europepmc.org/articles/PMC4572062.

[3] J. P. Bai and D. R. Abernethy. Systems pharmacology to predict drug toxicity: integration across levels of biological organization. *Annu. Rev. Pharmacol. Toxicol.*, 53: 451–473, 2013.

[4] D. P. Williams and B. K. Park. Idiosyncratic toxicity: the role of toxicophores and bioactivation. *Drug Discov. Today*, 8(22):1044–1050, Nov 2003.

[5] C. Calitz, L. du Plessis, C. Gouws, D. Steyn, J. Steenekamp, C. Muller, and S. Hamman. Herbal hepatotoxicity: current status, examples, and challenges. *Expert Opin Drug Metab Toxicol*, 11(10):1551–1565, 2015.

[6] D. M. Dambach, D. Misner, M. Brock, A. Fullerton, W. Proctor, J. Maher, D. Lee, K. Ford, and D. Diaz. Safety Lead Optimization and Candidate Identification: Integrating New Technologies into Decision-Making. *Chem. Res. Toxicol.*, Dec 2015.

[7] Cédric Merlot. Computational toxicology—a tool for early safety evaluation. *Drug Discovery Today*, 15(1–2):16 – 22, 2010. ISSN 1359-6446. doi: http://dx.doi.org/10.1016/j.drudis.2009.09.010. URL http://www.sciencedirect.com/science/article/pii/S1359644609003353.

[8] Nicholas J. Hrib and Norton P. Peet. Chemoinformatics: are we exploiting this new science?: 'we need to make chemoinformatics tools more accessible to the bench

chemist...'. *Drug Discovery Today*, 5(11):483 – 485, 2000. ISSN 1359-6446. doi: http://dx.doi.org/10.1016/S1359-6446(00)01560-9. URL http://www.sciencedirect.com/science/article/pii/S1359644600015609.

[9] Mike Hann and Richard Green. Chemoinformatics — a new name for an old problem? *Current Opinion in Chemical Biology*, 3(4):379 – 383, 1999. ISSN 1367-5931. doi: http://dx.doi.org/10.1016/S1367-5931(99)80057-X. URL http://www.sciencedirect.com/science/article/pii/S136759319980057X.

[10] Svava Ósk Jónsdóttir, Flemming Steen Jørgensen, and Søren Brunak. Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics*, 21(10):2145–2160, 2005. doi: 10.1093/bioinformatics/bti314. URL http://bioinformatics.oxfordjournals.org/content/21/10/2145.abstract.

[11] James F Blake. Chemoinformatics – predicting the physicochemical properties of 'drug-like' molecules. *Current Opinion in Biotechnology*, 11(1):104 – 107, 2000. ISSN 0958-1669. doi: http://dx.doi.org/10.1016/S0958-1669(99)00062-2. URL http://www.sciencedirect.com/science/article/pii/S0958166999000622.

[12] Chemical graphs, molecular matrices and topological indices in chemoinformatics and quantitative structure-activity relationships. .

[13] J. W. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, 16(7):521–533, Jul 2002.

[14] M. Karthikeyan* and Andreas Bender. Encoding and decoding graphical chemical structures as two-dimensional (pdf417) barcodes. *Journal of Chemical Information and Modeling*, 45(3):572–580, 2005. doi: 10.1021/ci049758i. URL http://dx.doi.org/10.1021/ci049758i.

[15] J. M. Barnard, G. M. Downs, A. von Scholley-Pfab, and R. D. Brown. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J. Mol. Graph. Model.*, 18(4-5):452–463, 2000.

[16] Wolfgang Guba, Agnes Meyder, Matthias Rarey, and Jérôme Hert. Torsion library reloaded: A new version of expert-derived smarts rules for assessing conformations of

small molecules. *Journal of Chemical Information and Modeling*, 56(1):1–5, 2016. doi: 10.1021/acs.jcim.5b00522. URL http://dx.doi.org/10.1021/acs.jcim.5b00522.

[17] Elisabet Gregori-Puigjané, Rut Garriga-Sust, and Jordi Mestres. Indexing molecules with chemical graph identifiers. *Journal of Computational Chemistry*, 32(12):2638–2646, 2011. ISSN 1096-987X. doi: 10.1002/jcc.21843. URL http://dx.doi.org/10.1002/jcc.21843.

[18] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1):1–9, 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-7. URL http://dx.doi.org/10.1186/1758-2946-5-7.

[19] Christopher Southan and Andras Stracz. Extracting and connecting chemical structures from text sources using chemicalize.org. *Journal of Cheminformatics*, 5(1):1–10, 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-20. URL http://dx.doi.org/10.1186/1758-2946-5-20.

[20] John R. Owen, Ian T. Nabney, José L. Medina-Franco, and Fabian López-Vallejo. Visualization of molecular fingerprints. *Journal of Chemical Information and Modeling*, 51(7):1552–1563, 2011. doi: 10.1021/ci1004042. URL http://dx.doi.org/10.1021/ci1004042.

[21] † Ling Xue, † Jeffrey W. Godden, † Florence L. Stahura, , and ‡ Jürgen Bajorath*, †. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *Journal of Chemical Information and Computer Sciences*, 43(4):1151–1157, 2003. doi: 10.1021/ci030285+. URL http://dx.doi.org/10.1021/ci030285+.

[22] Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, 11(2):137–148, 2016. doi: 10.1517/17460441.2016.1117070. URL http://dx.doi.org/10.1517/17460441.2016.1117070.

[23] Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. .

[24] † Jun Feng, ‡ Laura Lurati, § Haojun Ouyang, ‖ Tracy Robinson, Yuanyuan Wang, Shenglan Yuan, , and S. Stanley Young*. Predictive toxicology: benchmarking molecular descriptors and statistical methods. *Journal of Chemical Information and Computer Sciences*, 43(5):1463–1470, 2003. doi: 10.1021/ci034032s. URL http://dx.doi.org/10.1021/ci034032s.

[25] P. R. Andrews, D. J. Craik, and J. L. Martin. Functional group contributions to drug-receptor interactions. *Journal of Medicinal Chemistry*, 27(12):1648–1657, 1984. doi: 10.1021/jm00378a021. URL http://dx.doi.org/10.1021/jm00378a021.

[26] Zhisong He, Jian Zhang, Xiao-He Shi, Le-Le Hu, Xiangyin Kong, Yu-Dong Cai, and Kuo-Chen Chou. Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE*, 5:e9603, 03 2010. URL http://dx.doi.org/10.13712Fjournal.pone.0009603.

[27] R. T Sanderson. Models for demonstrating electronegativity and "partial charge". *Journal of Chemical Education*, 36(10):507, 1959.

[28] James Sangster. Octanolwater partition coefficients of simple organic compounds. *Journal of Physical and Chemical Reference Data*, 18(3), 1989.

[29] Sheng-Yong Yang. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today*, 15(11–12):444 – 450, 2010. ISSN 1359-6446. doi: http://dx.doi.org/10.1016/j.drudis.2010.03.013. URL http://www.sciencedirect.com/science/article/pii/S135964461000111X.

[30] John B. O. Mitchell. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(5):468–481, 2014. ISSN 1759-0884. doi: 10.1002/wcms.1183. URL http://dx.doi.org/10.1002/wcms.1183.

[31] Vinícius Gonçalves Maltarollo, Jadson Castro Gertrudes, Patrícia Rufino Oliveira, and Kathia Maria Honorio. Applying machine learning techniques for adme-tox prediction: a review. *Expert Opinion on Drug Metabolism & Toxicology*, 11(2):259–271, 2015. doi: 10.1517/17425255.2015.980814. URL http://dx.doi.org/10.1517/17425255.2015.980814.

[32] K. Kandasamy, J. K. Chuah, R. Su, P. Huang, K. G. Eng, S. Xiong, Y. Li, C. S. Chia, L. H. Loo, and D. Zink. Prediction of drug-induced nephrotoxicity and injury mechanisms

with human induced pluripotent stem cell-derived cells and machine learning methods. *Sci Rep*, 5:12337, 2015.

[33] Raja S Settivari, Nicholas Ball, Lynea Murphy, Reza Rasoulpour, Darrell R Boverhof, and Edward W Carney. Predicting the future: opportunities and challenges for the chemical industry to apply 21st-century toxicity testing. *Journal of the American Association for Laboratory Animal Science : JAALAS*, 54(2):214—223, March 2015. ISSN 1559-6109. URL http://europepmc.org/articles/PMC4382627.

[34] Alessandra Roncaglioni, Andrey A Toropov, Alla P Toropova, and Emilio Benfenati. In silico methods to predict drug toxicity. *Current Opinion in Pharmacology*, 13(5):802 – 806, 2013. ISSN 1471-4892. doi: http://dx.doi.org/10.1016/j.coph.2013.06.001. URL http://www.sciencedirect.com/science/article/pii/S1471489213000799. Anti-infectives • New technologies.

[35] I. Rusyn and G. P. Daston. Computational toxicology: realizing the promise of the toxicity testing in the 21st century. *Environ. Health Perspect.*, 118(8):1047–1050, Aug 2010.

[36] Malgorzata N Drwal, Priyanka Banerjee, Mathias Dunkel, Martin R Wettig, and Robert Preissner. Protox: a web server for the in silico prediction of rodent oral toxicity. *Nucleic acids research*, 42(Web Server issue):W53—8, July 2014. ISSN 0305-1048. doi: 10.1093/nar/gku401. URL http://europepmc.org/articles/PMC4086068.

[37] Robert P. Sheridan and Simon K. Kearsley. Why do we need so many chemical similarity search methods? *Drug Discovery Today*, 7(17):903 – 911, 2002. ISSN 1359-6446. doi: http://dx.doi.org/10.1016/S1359-6446(02)02411-X. URL http://www.sciencedirect.com/science/article/pii/S135964460202411X.

[38] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, and L. E. Weinberger. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.*, 39(16):3049–3059, Aug 1996.

[39] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, 25(2):197–206, Feb 2007.

[40] U. Schmidt, S. Struck, B. Gruening, J. Hossbach, I. S. Jaeger, R. Parol, U. Lindequist, E. Teuscher, and R. Preissner. SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.*, 37(Database issue):D295–299, Jan 2009.

[41] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL http://dx.doi.org/10.1021/ci100050t.

[42] Thomas Engel. *Representation of Chemical Compounds*, pages 15–168. Wiley-VCH Verlag GmbH Co. KGaA, 2004. ISBN 9783527601646. doi: 10.1002/3527601643.ch2. URL http://dx.doi.org/10.1002/3527601643.ch2.

[43] F. Eduati, L. M. Mangravite, T. Wang, H. Tang, J. C. Bare, R. Huang, T. Norman, M. Kellen, M. P. Menden, J. Yang, X. Zhan, R. Zhong, G. Xiao, M. Xia, N. Abdo, O. Kosyk, S. Friend, A. Dearry, A. Simeonov, R. R. Tice, I. Rusyn, F. A. Wright, G. Stolovitzky, Y. Xie, and J. Saez-Rodriguez. Erratum: Prediction of human population responses to toxic compounds by a collaborative competition. *Nat. Biotechnol.*, 33(10):1109, Oct 2015.

[44] David J. Rogers and Taffee T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960. ISSN 0036-8075. doi: 10.1126/science.132.3434.1115. URL http://science.sciencemag.org/content/132/3434/1115.

[45] Jessica Ahmed, Catherine L Worth, Paul Thaben, Christian Matzig, Corinna Blasse, Mathias Dunkel, and Robert Preissner. Fragmentstore–a comprehensive database of fragments linking metabolites, toxic molecules and drugs. *Nucleic acids research*, 39 (Database issue):D1049—54, January 2011. ISSN 0305-1048. doi: 10.1093/nar/gkq969. URL http://europepmc.org/articles/PMC3013803.

[46] Christian Hertweck. Natural products as source of therapeutics against parasitic diseases. *Angewandte Chemie International Edition*, 54(49):14622–14624, 2015. ISSN 1521-3773. doi: 10.1002/anie.201509828. URL http://dx.doi.org/10.1002/anie.201509828.

[47] .

[48] E. M. Driggers, S. P. Hale, J. Lee, and N. K. Terrett. The exploration of macrocycles for drug discovery–an underexploited structural class. *Nat Rev Drug Discov*, 7(7):608–624, Jul 2008.

[49] LEROY F. LIU, SHYAMAL D. DESAI, TSAI-KUN LI, YONG MAO, MEI SUN, and SAI-PENG SIM. Mechanism of action of camptothecin. *Annals of the New York Academy of Sciences*, 922(1):1–10, 2000. ISSN 1749-6632. doi: 10.1111/j.1749-6632.2000.tb07020.x. URL http://dx.doi.org/10.1111/j.1749-6632.2000.tb07020.x.

[50] Yves Pommier, Juana M. Barcelo, V. Ashutosh Rao, Olivier Sordet, Andrew G. Jobson, Laurent Thibaut, ZeHong Miao, Jennifer A. Seiler, Hongliang Zhang, Christophe Marchand, Keli Agama, John L. Nitiss, and Christophe Redon. Repair of topoisomerase imediated {DNA} damage. volume 81 of *Progress in Nucleic Acid Research and Molecular Biology*, pages 179 – 229. Academic Press, 2006. doi: http://dx.doi.org/10.1016/S0079-6603(06)81005-6. URL http://www.sciencedirect.com/science/article/pii/S0079660306810056.

[51] T. Ferenc, B. Lukasiewicz, J. Cie?wierz, and E. Kowalczyk. [Poisonings with Amanita phalloides]. *Med Pr*, 60(5):415–426, 2009.

[52] Barry V. Charlwood. Chcd dictionary of natural products on cd-rom. chapman and hall, london, uk, 1992. £2,950 first year's subscription; £1,750 annual renewal—academic discounts available. issn (cd-rom) 0966 2146. *Phytochemical Analysis*, 4(3):135–137, 1993. ISSN 1099-1565. doi: 10.1002/pca.2800040311. URL http://dx.doi.org/10.1002/pca.2800040311.

[53] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27. URL http://nar.oxfordjournals.org/content/28/1/27.abstract.

[54] Ali Abdel-Rahman, Njwen Anyangwe, Louis Carlacci, Steve Casper, Rebecca P. Danam, Evaristus Enongene, Gladys Erives, Daniel Fabricant, Ramadevi Gudi, Corey J. Hilmas, Fred Hines, Paul Howard, Dan Levy, Ying Lin, Robert J. Moore, Erika Pfeiler, T. Scott Thurmond, Saleh Turujman, and Nigel J. Walker. The safety and regulation of natural products used as foods and food ingredients. *Toxicological Sciences*, 123(2):333–348, 2011. doi: 10.1093/toxsci/kfr198. URL http://toxsci.oxfordjournals.org/content/123/2/333.abstract.

[55] M. Nauffal and S. Gabardi. Nephrotoxicity of Natural Products. *Blood Purif.*, 41(1-3):123–129, Jan 2016.

[56] Mohammadreza Ardalan, Zahra Samadifar, and Amir Vahedi. Creatine monohydrate supplement induced interstitial nephritis. *Journal of Nephropathology*, 1. ISSN 2251-8363. doi: 10.5812/nephropathol.7530. URL http://nephropathol.com/Abstract/JNP_20130216141036.

[57] Malgorzata N. Drwal, Priyanka Banerjee, Mathias Dunkel, Martin R. Wettig, and Robert Preissner. Protox: a web server for the in silico prediction of rodent oral toxicity. *Nucleic Acids Research*, 42(W1):W53–W58, 2014. doi: 10.1093/nar/gku401. URL http://nar.oxfordjournals.org/content/42/W1/W53.abstract.

[58] Evan E. Bolton, Yanli Wang, Paul A. Thiessen, and Stephen H. Bryant. Chapter 12 - pubchem: Integrated platform of small molecules and biological activities. volume 4 of *Annual Reports in Computational Chemistry*, pages 217 – 241. Elsevier, 2008. doi: http://dx.doi.org/10.1016/S1574-1400(08)00012-1. URL http://www.sciencedirect.com/science/article/pii/S1574140008000121.

[59] Stefan Günther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen, Reinhard Schneider, Roman Skoblo, Robert B. Russell, Philip E. Bourne, Peer Bork, and Robert Preissner. Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Research*, 36(suppl 1):D919–D922, 2008. doi: 10.1093/nar/gkm862. URL http://nar.oxfordjournals.org/content/36/suppl_1/D919.abstract.

[60] Janette Nickel, Bjoern-Oliver Gohlke, Jevgeni Erehman, Priyanka Banerjee, Wen Wei Rong, Andrean Goede, Mathias Dunkel, and Robert Preissner. Superpred: update on drug classification and target prediction. *Nucleic Acids Research*, 2014. doi: 10.1093/nar/gku477. URL http://nar.oxfordjournals.org/content/early/2014/05/30/nar.gku477.abstract.

[61] Dominik M. Peter, Lennart SchadavonBorzyskowski, Patrick Kiefer, Philipp Christen, Julia A. Vorholt, and Tobias J. Erb. Screening and engineering the synthetic potential of carboxylating reductases from central metabolism and polyketide biosynthesis. *Angewandte Chemie International Edition*, 54(45):13457–13461, 2015. ISSN 1521-3773. doi: 10.1002/anie.201505282. URL http://dx.doi.org/10.1002/anie.201505282.

[62] Shagun Krishna, Vikash Kumar, and Mohammad Imran Siddiqi. Recent advances in computer-assisted structure-based identification and design of histone deacetylases inhibitors. *Current Topics in Medicinal Chemistry*, 16(9), 2016-04-01T00:00:00. URL http://www.ingentaconnect.com/content/ben/ctmc/2016/00000016/00000009/art00007.

[63] Vishal B. Siramshetty, Janette Nickel, Christian Omieczynski, Bjoern-Oliver Gohlke, Malgorzata N. Drwal, and Robert Preissner. Withdrawn—a resource for withdrawn and discontinued drugs. *Nucleic Acids Research*, 44(D1):D1080–D1086, 2016. doi: 10.1093/nar/gkv1192. URL http://nar.oxfordjournals.org/content/44/D1/D1080.abstract.

[64] Malgorzata Natalia Drwal, Vishal Babu Siramshetty, Priyanka Banerjee, Andrean Goede, Robert Preissner, and Mathias Dunkel. Molecular similarity-based predictions of the tox21 screening outcome. *Frontiers in Environmental Science*, 3(54), 2015. ISSN 2296-665X. doi: 10.3389/fenvs.2015.00054. URL http://www.frontiersin.org/environmental_informatics/10.3389/fenvs.2015.00054/abstract.

[65] Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Anna Rossoshek, and Anton Simeonov. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3(85), 2016. ISSN 2296-665X. doi: 10.3389/fenvs.2015.00085. URL http://www.frontiersin.org/environmental_informatics/10.3389/fenvs.2015.00085/abstract.

[66] Sereina Riniker and Gregory A. Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):1–17, 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-26. URL http://dx.doi.org/10.1186/1758-2946-5-26.

[67] Lowell H. Hall and Lemont B. Kier. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *Journal of Chemical Information and Computer Sciences*, 35(6):1039–1045, 1995. doi: 10.1021/ci00028a014. URL http://dx.doi.org/10.1021/ci00028a014.

[68] Stephan Beisken, Thorsten Meinl, Bernd Wiswedel, Luis F de Figueiredo, Michael Berthold, and Christoph Steinbeck. Knime-cdk: Workflow-driven cheminformatics. *BMC bioinformatics*, 14:257, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-257. URL http://europepmc.org/articles/PMC3765822.

[69] T. J. Hou, , and X. J. Xu*. Adme evaluation in drug discovery. 3. modeling blood-brain barrier partitioning using simple molecular descriptors. *Journal of Chemical Information and Computer Sciences*, 43(6):2137–2152, 2003. doi: 10.1021/ci034134i. URL http://dx.doi.org/10.1021/ci034134i.

[70] Lowell H. Hall and Lemont B. Kier. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling*, pages 367–422. John Wiley Sons, Inc., 2007. ISBN 9780470125793. doi: 10.1002/9780470125793.ch9. URL http://dx.doi.org/10.1002/9780470125793.ch9.

[71] † Nidhi, ‡ Meir Glick, ‡ John W. Davies, , and ‡ Jeremy L. Jenkins*. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *Journal of Chemical Information and Modeling*, 46 (3):1124–1133, 2006. doi: 10.1021/ci060003g. URL http://dx.doi.org/10.1021/ci060003g.

[72] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.

[73] Samuel J. Webb, Thierry Hanser, Brendan Howlin, Paul Krause, and Jonathan D. Vessey. Feature combination networks for the interpretation of statistical machine learning models: application to ames mutagenicity. *Journal of Cheminformatics*, 6(1):1–21, 2014. ISSN 1758-2946. doi: 10.1186/1758-2946-6-8. URL http://dx.doi.org/10.1186/1758-2946-6-8.

[74] Anthony Zaknich. The multiclass probabilistic neural network (pnn) classifier. Technical report, 1990.

[75] J. Addeh, A. Ebrahimzadeh, M. Azarbad, and V. Ranaee. Statistical process control using optimized neural networks: a case study. *ISA Trans*, 53(5):1489–1499, Sep 2014.

[76] Q. Zhang. Dynamic Uncertain Causality Graph for Knowledge Representation and Probabilistic Reasoning: Directed Cyclic Graph and Joint Probability Distribution. *IEEE Trans Neural Netw Learn Syst*, 26(7):1503–1517, Jul 2015.

[77] M H Zweig and G Campbell. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–77, 1993. URL http://www.clinchem.org/content/39/4/561.abstract.

[78] Y. A. Skaik. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*, 56(4):341; author reply 341, 2008.

[79] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. Deeptox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3(80), 2016. ISSN 2296-665X. doi: 10.3389/fenvs.2015.00080. URL http://www.frontiersin.org/environmental_informatics/10.3389/fenvs.2015.00080/abstract.

[80] Yoshihiro Uesawa. Rigorous selection of random forest models for identifying compounds that activate toxicity-related pathways. *Frontiers in Environmental Science*, 4(9), 2016. ISSN 2296-665X. doi: 10.3389/fenvs.2016.00009. URL http://www.frontiersin.org/environmental_informatics/10.3389/fenvs.2016.00009/abstract.

[81] A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, M. Uhlen, and J. Nielsen. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat Commun*, 5:3083, 2014.

[82] Christoph Gille, Christian Bölling, Andreas Hoppe, Sascha Bulik, Sabrina Hoffmann, Katrin Hübner, Anja Karlstädt, Ramanan Ganeshan, Matthias König, Kristian Rother, Michael Weidlich, Jörn Behre, and Herrmann-Georg Holzhütter. Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Molecular Systems Biology*, 6(1), 2010. doi: 10.1038/msb.2010.62. URL http://msb.embopress.org/content/6/1/411.

[83] Dany Habka, David Mann, Ronald Landes, and Alejandro Soto-Gutierrez. Future economics of liver transplantation: A 20-year cost modeling forecast and the prospect of bioengineering autologous liver grafts. *PLoS ONE*, 10(7):1–21, 07 2015. doi: 10.1371/journal.pone.0131764. URL http://dx.doi.org/10.1371%2Fjournal.pone.0131764.

[84] Sonia Roman and Arturo Panduro. Genomic medicine in gastroenterology: A new approach or a new specialty? *World journal of gastroenterology*, 21(27):8227—8237, July 2015. ISSN 1007-9327. doi: 10.3748/wjg.v21.i27.8227. URL http://europepmc.org/articles/PMC4507092.

[85] Liane Rabinowich and Oren Shibolet. Drug induced steatohepatitis: An uncommon culprit of a common disease. *BioMed research international*, 2015:168905, 2015. ISSN 2314-6133. doi: 10.1155/2015/168905. URL http://europepmc.org/articles/PMC4529891.

[86] Paola Dongiovanni, Raffaela Rametta, Marica Meroni, and Luca Valenti. The role of insulin resistance in nonalcoholic steatohepatitis and liver disease development - a potential therapeutic target? *Expert review of gastroenterology hepatology*, 10(2): 229—242, February 2016. ISSN 1747-4124. doi: 10.1586/17474124.2016.1110018. URL http://dx.doi.org/10.1586/17474124.2016.1110018.

[87] Ildefonso Martínez De La Fuente. Elements of the cellular metabolic structure. *Frontiers in Molecular Biosciences*, 2(16), 2015. ISSN 2296-889X. doi: 10.3389/fmolb.2015.00016. URL http://www.frontiersin.org/mathematics_of_biomolecules/10.3389/fmolb.2015.00016/abstract.

[88] M. Stitt, R. Sulpice, and J. Keurentjes. Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant Physiol.*, 152(2):428–444, Feb 2010.

[89] Jing Tang and Tero Aittokallio. Network pharmacology strategies toward multi-target anticancer therapies: from computational models to experimental design principles. *Current pharmaceutical design*, 20(1):23—36, 2014. ISSN 1381-6128. doi: 10.2174/13816128113199990470. URL http://dx.doi.org/10.2174/13816128113199990470.

[90] Ralf Steuer, Thilo Gross, Joachim Selbig, and Bernd Blasius. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences*, 103 (32):11868–11873, 2006. doi: 10.1073/pnas.0600013103. URL http://www.pnas.org/content/103/32/11868.abstract.

[91] J. T. Dean, M. L. Rizk, Y. Tan, K. M. Dipple, and J. C. Liao. Ensemble modeling of hepatic fatty acid metabolism with a synthetic glyoxylate shunt. *Biophys. J.*, 98(8):1385–1395, Apr 2010.

[92] Liang Tong and H. James Harwood. Acetyl-coenzyme a carboxylases: Versatile targets for drug discovery. *Journal of Cellular Biochemistry*, 99(6):1476–1488, 2006. ISSN 1097-4644. doi: 10.1002/jcb.21077. URL http://dx.doi.org/10.1002/jcb.21077.

[93] Matthew P. Bourbeau and Michael D. Bartberger. Recent advances in the development of acetyl-coa carboxylase (acc) inhibitors for the treatment of metabolic disease. *Journal of Medicinal Chemistry*, 58(2):525–536, 2015. doi: 10.1021/jm500695e. URL http://dx.doi.org/10.1021/jm500695e.

[94] Victor A. Zammit. Carnitine palmitoyltransferase 1: Central to cell function. *IUBMB Life*, 60(5):347–354, 2008. ISSN 1521-6551. doi: 10.1002/iub.78. URL http://dx.doi.org/10.1002/iub.78.

[95] Simona M. Ceccarelli, Odile Chomienne, Marcel Gubler, and Arduino Arduini. Carnitine palmitoyltransferase (cpt) modulators: A medicinal chemistry perspective on 35 years of research. *Journal of Medicinal Chemistry*, 54(9):3109–3152, 2011. doi: 10.1021/jm100809g. URL http://dx.doi.org/10.1021/jm100809g.

[96] Albert Lehninger, David L. Nelson, and Michael M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, fifth edition edition, June 2008. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20%255C&path=ASIN/1429224169.

[97] Donghai Wu, Lakshmanan Govindasamy, Wei Lian, Yunrong Gu, Thomas Kukar, Mavis Agbandje-McKenna, and Robert McKenna. Structure of human carnitine acetyltransferase: Molecular basis for fatty acyl transfer. *Journal of Biological Chemistry*, 278 (15):13159–13165, 2003. doi: 10.1074/jbc.M212356200. URL http://www.jbc.org/content/278/15/13159.abstract.

[98] Eduardo López-Viñas, Assia Bentebibel, Chandrashekaran Gurunathan, Montserrat Morillas, Dolores de Arriaga, Dolors Serra, Guillermina Asins, Fausto G. Hegardt, and Paulino Gómez-Puertas. Definition by functional and structural analysis of two malonyl-coa sites in carnitine palmitoyltransferase 1a. *Journal of Biological Chemistry*, 282(25):18212–18224, 2007. doi: 10.1074/jbc.M700885200. URL http://www.jbc.org/content/282/25/18212.abstract.

[99] Jean-Paul Bonnefont, Fatima Djouadi, Carina Prip-Buus, Stephanie Gobin, Arnold Munnich, and Jean Bastin. Carnitine palmitoyltransferases 1 and 2: biochemical, molecular and medical aspects. *Molecular Aspects of Medicine*, 25(5–6):495 – 520, 2004. ISSN 0098-2997. doi: http://dx.doi.org/10.1016/j.mam.2004.06.004. URL http:

//www.sciencedirect.com/science/article/pii/S0098299704000494. Carnitine.

[100] Yu-Shan Hsiao, Gerwald Jogl, and Liang Tong. Crystal structures of murine carnitine acetyltransferase in ternary complexes with its substrates. *Journal of Biological Chemistry*, 281(38):28480–28487, 2006. doi: 10.1074/jbc.M602622200. URL http://www.jbc.org/content/281/38/28480.abstract.

[101] Karima Begriche, Julie Massart, Marie-Anne Robin, Annie Borgne-Sanchez, and Bernard Fromenty. Drug-induced toxicity on mitochondria and lipid metabolism: Mechanistic diversity and deleterious consequences for the liver. *Journal of Hepatology*, 54(4):773 – 794, 2011. ISSN 0168-8278. doi: http://dx.doi.org/10.1016/j.jhep.2010.11.006. URL http://www.sciencedirect.com/science/article/pii/S0168827810010664.

[102] Montserrat MORILLAS, Eduardo LÓPEZ-VIÑAS, Alfonso VALENCIA, Dolors SERRA, Paulino GÓMEZ-PUERTAS, Fausto G. HEGARDT, and Guillermina ASINS. Structural model of carnitine palmitoyltransferase i based on the carnitine acetyltransferase crystal. 379(3):777–784, 2004. doi: 10.1042/bj20031373. URL http://www.biochemj.org/content/379/3/777.

[103] Hans-Christian Ehrlich and Matthias Rarey. Systematic benchmark of substructure search in molecular graphs - from ullmann to vf2. *Journal of Cheminformatics*, 4(1):1–17, 2012. ISSN 1758-2946. doi: 10.1186/1758-2946-4-13. URL http://dx.doi.org/10.1186/1758-2946-4-13.

[104] Munir Pirmohamed, Sally James, Shaun Meakin, Chris Green, Andrew K Scott, Thomas J Walley, Keith Farrar, B Kevin Park, and Alasdair M Breckenridge. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ (Clinical research ed.)*, 329(7456):15—19, July 2004. ISSN 0959-8138. doi: 10.1136/bmj.329.7456.15. URL http://europepmc.org/articles/PMC443443.

[105] Jack Uetrecht and Dean J. Naisbitt. Idiosyncratic adverse drug reactions: Current concepts. 65(2):779–808, 2013. doi: 10.1124/pr.113.007450. URL http://pharmrev.aspetjournals.org/content/65/2/779.abstract.

[106] P. T. Illing, J. P. Vivian, N. L. Dudek, L. Kostenko, Z. Chen, M. Bharadwaj, J. J. Miles, L. Kjer-Nielsen, S. Gras, N. A. Williamson, S. R. Burrows, A. W. Purcell, J. Rossjohn, and

J. McCluskey. Immune self-reactivity triggered by drug-modified HLA-peptide repertoire. *Nature*, 486(7404):554–558, Jun 2012.

[107] David A. Ostrov, Barry J. Grant, Yuri A. Pompeu, John Sidney, Mikkel Harndahl, Scott Southwood, Carla Oseroff, Shun Lu, Jean Jakoncic, Cesar Augusto F. de Oliveira, Lun Yang, Hu Mei, Leming Shi, Jeffrey Shabanowitz, A. Michelle English, Amanda Wriston, Andrew Lucas, Elizabeth Phillips, Simon Mallal, Howard M. Grey, Alessandro Sette, Donald F. Hunt, Soren Buus, and Bjoern Peters. Drug hypersensitivity caused by alteration of the mhc-presented self-peptide repertoire. *Proceedings of the National Academy of Sciences*, 109(25):9959–9964, 2012. doi: 10.1073/pnas.1207934109.

[108] T. Profaizer and D. Eckels. Hla alleles and drug hypersensitivity reactions. *International Journal of Immunogenetics*, 39(2):99–105, 2012. ISSN 1744-313X. doi: 10.1111/j.1744-313X.2011.01061.x. URL http://dx.doi.org/10.1111/j.1744-313X.2011.01061.x.

[109] Diana Chessman, Lyudmila Kostenko, Tessa Lethborg, Anthony W. Purcell, Nicholas A. Williamson, Zhenjun Chen, Lars Kjer-Nielsen, Nicole A. Mifsud, Brian D Tait, Rhonda Holdsworth, Coral Ann Almeida, David Nolan, Whitney A. Macdonald, Julia K. Archbold, Anthony D. Kellerher, Debbie Marriott, Simon Mallal, Mandvi Bharadwaj, Jamie Rossjohn, and James McCluskey. Human leukocyte antigen class i-restricted activation of cd8+ t cells provides the immunogenetic basis of a systemic drug hypersensitivity. *Immunity*, 28(6):822 – 832, 2008. ISSN 1074-7613. doi: http://dx.doi.org/10.1016/j.immuni.2008.04.020. URL http://www.sciencedirect.com/science/article/pii/S1074761308002422.

[110] M. S. Anderson, T. N. Kakuda, W. Hanley, J. Miller, J. T. Kost, R. Stoltz, L. A. Wenning, J. A. Stone, R. M. Hoetelmans, J. A. Wagner, and M. Iwamoto. Minimal pharmacokinetic interaction between the human immunodeficiency virus nonnucleoside reverse transcriptase inhibitor etravirine and the integrase inhibitor raltegravir in healthy subjects. *Antimicrob. Agents Chemother.*, 52(12):4228–4232, Dec 2008.

[111] S. N. Lavergne, B. K. Park, and D. J. Naisbitt. The roles of drug metabolism in the pathogenesis of T-cell-mediated drug hypersensitivity. *Curr Opin Allergy Clin Immunol*, 8(4):299–307, Aug 2008.

[112] W. J. Pichler. The p-i Concept: Pharmacological Interaction of Drugs With Immune Receptors. *World Allergy Organ J*, 1(6):96–102, Jun 2008.

[113] Munir Pirmohamed, Dean J Naisbitt, Fraser Gordon, and B.Kevin Park. The danger hypothesis—potential role in idiosyncratic drug reactions. *Toxicology*, 181–182:55 – 63, 2002. ISSN 0300-483X. doi: http://dx.doi.org/10.1016/S0300-483X(02)00255-X. URL http://www.sciencedirect.com/science/article/pii/S0300483X0200255X.

[114] R. A. Bauer, P. E. Bourne, A. Formella, C. Frommel, C. Gille, A. Goede, A. Guerler, A. Hoppe, E. W. Knapp, T. Poschel, B. Wittig, V. Ziegler, and R. Preissner. Superimpose: a 3D structural superposition server. *Nucleic Acids Res.*, 36(Web Server issue):47–54, Jul 2008.

[115] Imir G. Metushi, Amanda Wriston, Priyanka Banerjee, Bjoern Oliver Gohlke, A. Michelle English, Andrew Lucas, Carrie Moore, John Sidney, Soren Buus, David A. Ostrov, Simon Mallal, Elizabeth Phillips, Jeffrey Shabanowitz, Donald F. Hunt, Robert Preissner, and Bjoern Peters. Acyclovir has low but detectable influence on hla-b*57:01 specificity without inducing hypersensitivity. *PLoS ONE*, 10(5):e0124878, 05 2015. doi: 10.1371/journal.pone.0124878. URL http://dx.doi.org/10.1371%2Fjournal.pone.0124878.

[116] RA Laskowski, N Furnham, and JM Thornton. The ramachandran plot and protein structure validation. In *Biomolecular Forms and Functions: A Celebration of 50 Years of the Ramachandran Map*, page 62—75. World Scientific Publishing, 2013. ISBN 9789814449137. doi: 10.1142/9789814449144_0005. URL http://dx.doi.org/10.1142/9789814449144_0005.

Appendix A

Software and databases

Different open source as well as commercial softwares used in this thesis are acknowledged in this section.

Name	License	Original language	Links
Open Babel	Open	C++	http://openbabel.org/
JChem	Academic	Java	https://www.chemaxon.com
MyChem	Open	C++	http://mychem.sourceforge.net/
KNIME	Open	Java	https://www.knime.org/
Discovery Studio	Commercial	NA	http://accelrys.com/products/collaborative-science/biovia-discovery-studio/
GOLD	Commercial	NA	http://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/
PyMOL	Academic	NA	https://www.pymol.org/

Additionally, the databases used in the different projects reported in this thesis are mentioned.

Name	license	links
PubChem	Open	https://pubchem.ncbi.nlm.nih.gov/
ChEMBL	Open	https://www.ebi.ac.uk/chembl/
PDB	Open	http://www.rcsb.org/pdb/home/home.do
ZINC	Open	http://zinc.docking.org/
SuperToxic	Open	http://bioinformatics.charite.de/supertoxic/
DrugBank	Open	http://www.drugbank.ca/
PubMed	Open	http://www.ncbi.nlm.nih.gov/pubmed
LiverTox	Open	http://livertox.nih.gov/
KEGG	Open	http://www.genome.jp/kegg/

Appendix B

Ehrenwortliche Erklarung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, February 2016